A CONFIABILIDADE DO MODELO TRAFFIC GPT : ADAPTAÇÃO PARA UM CONTEXTO DE ATAQUES ZERO-DAY

Monografia II em Sistemas de Informação

Autora: Ana Luiza Macêdo (analuizamacedost@gmail.com) **Orientadora:** Michele Nogueira (michele@dcc.ufmg.br)

Introdução

Ataques zero-day exigem modelos capazes de reconhecer padrões inéditos em tráfego criptografado, onde técnicas tradicionais falham. O TrafficGPT(Ginige et al., 2024) utiliza o GPT-2 para interpretar esse tráfego como linguagem, mas sofre de superconfiança, podendo classificar erroneamente dados desconhecidos, algo crítico em segurança. Este trabalho adapta o modelo para cenários open-set, aplicando calibração para reduzir a superconfiança e tornar as decisões mais confiáveis, permitindo que o TrafficGPT apoie de forma segura sistemas de rede como SIEM, IDS e firewalls.

Objetivos

O objetivo geral deste trabalho é aumentar a confiabilidade da classificação open-set do TrafficGPT, reduzindo a superconfiança do GPT-2 ao lidar com tráfego desconhecido. Para isso, buscamos otimizar o desempenho do classificador k-LND ajustando a forma como ele absorve os logits do modelo, aplicando uma camada de calibração baseada em escalonamento de temperatura, na qual os logits são divididos por um fator T. Essa estratégia segue a lógica matemática do Softmax, que controla a inclinação das probabilidades, e utiliza o método L-BFGS-B para encontrar o valor ótimo de T, tornando a confiança do modelo mais honesta e robusta frente a cenários zero-day

Conceitos Fundamentais

- *Classificação Open-Set:* o modelo precisa reconhecer o que é conhecido e rejeitar padrões inéditos, evitando classificar ataques zero-day como tráfego legítimo

- *Os Logits e Softmax:*

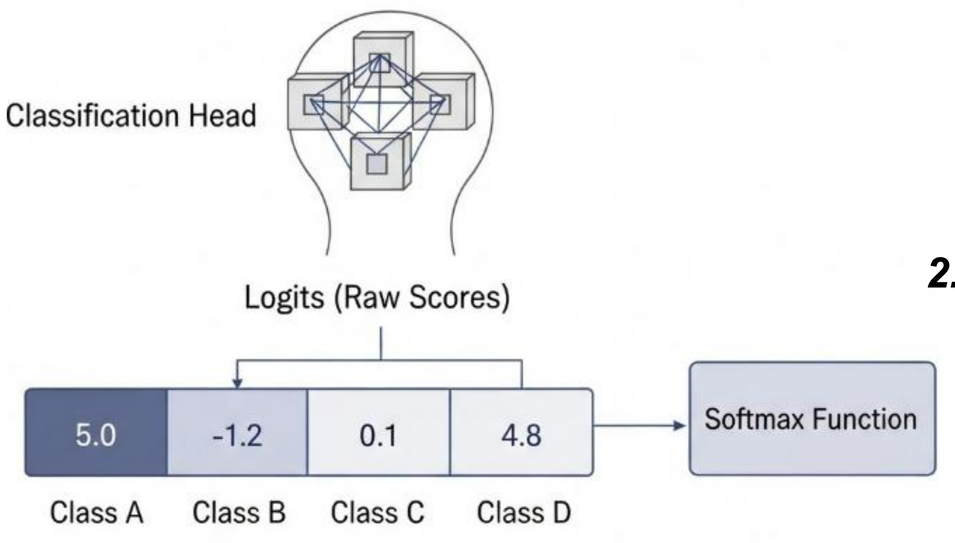

Figura 1. A arquitetura da Classificação

Metodologia

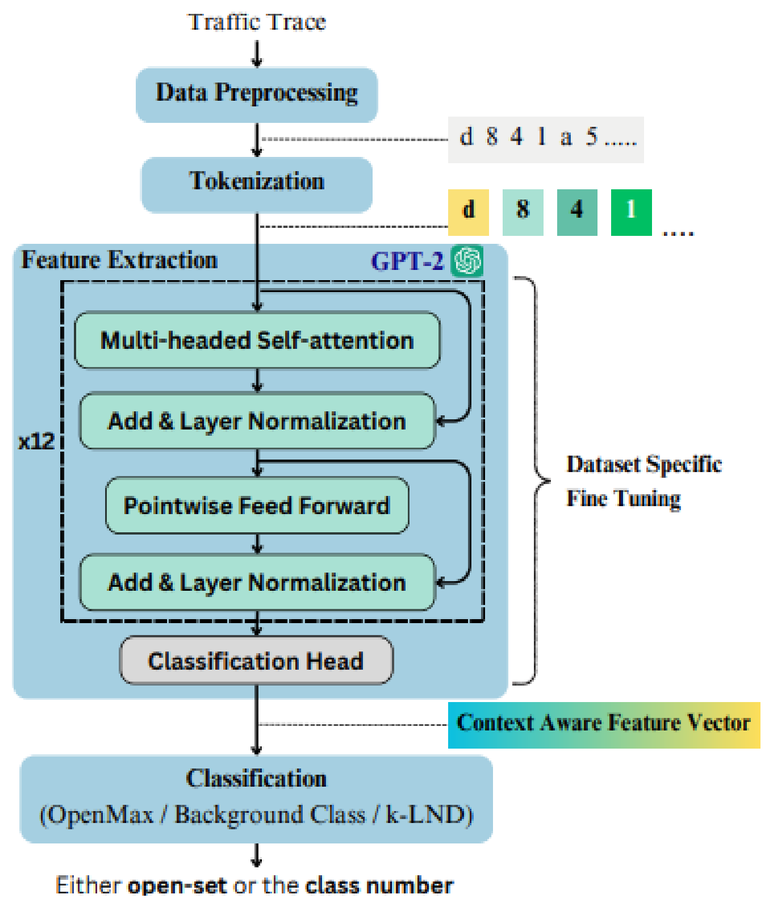

Figura 2. Arquitetura Original do Traffic GPT(Ginige et al.,2024)

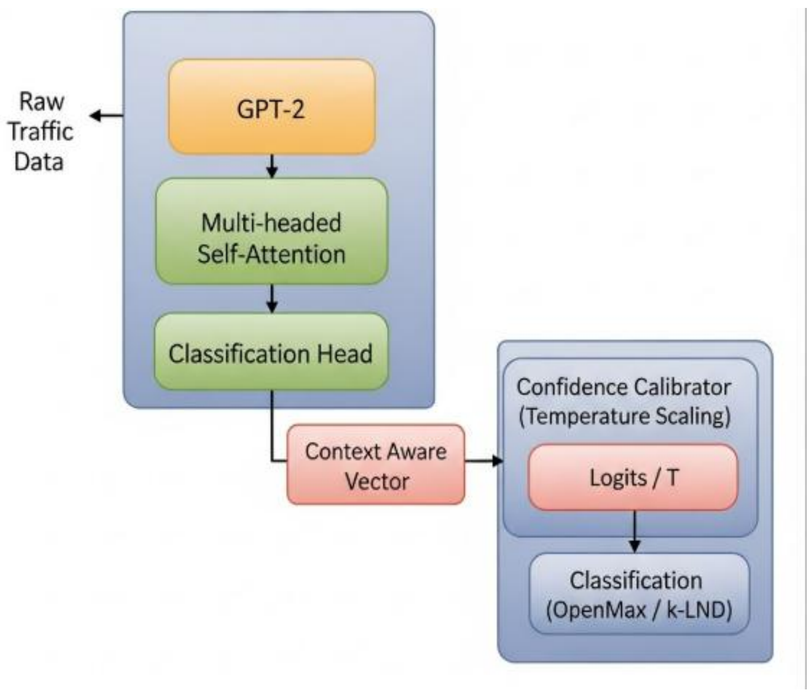

Figura 3. Nova Arquitetura do Traffic GPT

1. ***Calibração de Confiança do Modelo***: O pesquisador otimiza o parâmetro de Temperatura (T) mediante algoritmo L-BFGS-B para corrigir a superconfiança inerente do GPT-2 em cenários de dados inéditos, assegurando estimativas probabilísticas mais confiáveis.

2. ***Adaptação por Escalamento de Temperatura:*** Aplica-se Temperature Scaling como etapa de pós-processamento aos logits brutos do modelo , ajustando a confiança das previsões sem necessidade de retreinamento do modelo base.

3. ***Análise do impacto:*** inferências sobre a nova classificação do dataset CICIDS-2017, utilizado para a simulação do cenário Zero-Day, e comparar o T encontrado em relação ao dataset de treinamento (DC Dataset)

Resultados

Os resultados obtidos nos dois cenários experimentais revelam nuances importantes sobre calibração de confiança em modelos de deep learning para segurança. No DC Dataset, o modelo GPT-2 demonstrou calibração natural (T=1.0000) validando a arquitetura em condições controladas. Já no CIC-IDS2017, a superconfiança quantificada (T=1.4000) evidenciou o desafio real de zero-day, porém a relação direta entre calibração e melhoria métrica não foi completamente estabelecida, sugerindo que a temperatura sozinha não captura toda a complexidade do problema.

Conclusão

Os resultados obtidos indicam que modelos baseados em LLMs, como o TrafficGPT, possuem forte capacidade de representação e aprendem padrões complexos de tráfego cifrado mesmo em cenários fechados. Entretanto, diante de tráfego desconhecido esses modelos revelam uma tendência natural à superconfiança, o que compromete sua utilização direta em segurança. A calibração de temperatura demonstrou ser um mecanismo que pode indicar a necessidade de tornar as saídas do modelo mais honestas e, portanto, mais úteis em classificadores open-set como o k-LND.

Assim, o trabalho evidencia que LLMs podem, sim, compor sistemas de detecção de ameaças mais robustos, mas como componentes que requerem ajuste fino, análise estrutural e integração criteriosa com classificadores e módulos de decisão. A evolução desse tipo de abordagem depende, portanto, da exploração completa da arquitetura permitindo que modelos generativos sejam realmente confiáveis e eficazes em ambientes adversariais e dinâmicos como o de ataques zero-day.

https://dcc.ufmg.br

@dcc.ufmg

DCC
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO

UFMG