Título: Investigação de malware com técnicas BERT e Vision Transformers

1 Introdução

Com a evolução da tecnologia, o mundo atual encontra-se significativamente conectado via diversos dispositivos eletrônicos integrados à internet. O fluxo de informação que circula pela internet é extremamente volumoso, podendo conter qualquer tipo de dado, sendo este legítimo ou malicioso. Logo, "software maliciosos" (malware), códigos de software ou programa de computadores intencionalmente escrito para prejudicar um sistema de computador ou seus usuários, são constantemente aprimorados para difentes crimes cibernéticos, tais como: utilizar dispositivos, dados ou redes corporativas inteiras reféns para obter recursos financeiros; ter acesso não autorizado de dados confidenciais ou ativos digitais; obter credenciais de sistemas, números de cartão de crédito ou propriedades intelectuais; interromper sistemas críticos de empresas e/ou agências governamentais. A comunicação pela internet é estabelecida via pedidos e respostas, que seguem um protocolo, sendo o *hypertext transfer protocol* (HTTP) um dos mais comuns. Servidores HTTP estão abertos para qualquer tipo de requisição, que podem conter informações maliciosas e que caracterizam distintas ameaças. Também, é notável o crescente número de ataques cibernéticos registrados pela internet, com rápida propagação e visível sofisticação, seja por e-mails, links contaminados, aplicativos maliciosos ou *malwares*.

Métodos desenvolvidos para detectar malwares são extremamente necessários, com destaque para soluções fundamentadas em técnicas de inteligência artificial. Entretanto, como abordado no estudo de Xuejun Zong et al. [**?**], modelos obtidos via CNNs, RNNs e GANs possuem limitações que representam obstáculos significativos para uma aplicação prática efetiva, impactando o uso e a confiabilidade destes em contextos críticos. Por outro lado, o uso e abordagens *transformers* têm minimizado parte das dificuldades, com modelos capazes de indicar as maiores taxas de sucesso e com os melhores níveis de confiabilidade. Utilizando os *encoders*, componentes responsáveis por transformar as entradas (imagem ou texto) em números, que podem assim finalmente ser interpretados pelo modelo, os *transformers* se diferenciam das outras arquiteturas devido ao mecanismo de *attention*. Este consiste em realizar comparações com o conteúdo dos *datasets* utilizados no treinamento e examinar adicionalmente cada *token* relacionado aos outros *tokens* de entrada, provendo diversos valores semânticos para um mesmo dado. Este fato possibilita ao modelo compreender o contexto no qual um determinado *token* está inserido.

Neste contexto, o *Bidirectional Encoder Representations from Transformers* (BERT) e os *Vision Transformers* (ViT) merecem destaque em tarefas de classificação e reconhecimento de padrões de textos e imagens, respectivamente. Assim, considerando os avanços indicados previamente, é possível investigar o poder discriminativos dessas técnicas na área da *cybersegurança*, especificamente no campo de identificação de malware. BERT tem sido amplamente utilizado para o reconhecimento e classificação de tipos de malware, tais como registros de captura de pacotes do tráfego de rede, *SecurityBERT* [**?**], domínios maliciosos, código binário executável de programas, e até mesmo logs dos processos registrados por sistemas em dispositivos, como o CyBERT [**?**]. Por outro lado, modelo ViT é uma tecnologia ainda pouca explorada neste tipo de aplicação, mesmo com algumas abordagens relevantes na área de aplicação desta proposta, como B_ViT [**?**] e DE-ViT [**?**]. A aplicação de ViT neste contexto torna-se viável a partir da representação de texto em imagens. Portanto, esse tipo de estudo e investigação permite contribuições relevantes para a classificação e o reconhecimento de padrões na área de *cybersegurança*, por fornecerem testes ainda não explorados, especialmente via diferentes técnicas para representação de textos (dados que podem caracterizar cada tipo de malware) comumente em imagens. Os conhecimentos obtidos podem embasar o desenvolvimento de sistemas computacionais mais completos e confiáveis, além de observar possíveis limites no contexto de ViT para essa finalidade.

2 Objetivos

Neste projeto, a meta é investigar e aplicar modelos fundamentados em *transformers* para classificar e reconhecer padrões na área de *CyberAtaques*, tais como via requisições HTTP/HTTPS. Para tanto, pretende-se:

- Explorar um modelo BERT com uma abordagem *multilayer perceptron* (MLP) para classificar e reconhecer os padrões sob investigação;
- Investigar técnicas para transformação de texto em imagens (*reshaping*) a fim de testar a viabilidade ViT neste tipo aplicação (*CyberAtaques*), tais como amostras de *distributed denial of service* (DDoS), *man in the middle* (MitM), SQL injection ou *cross site scripting* (XSS) [**?**];
- Explorar o modelo ViT nomeado como *Butterfly Vision Transformer* para classificar as representações obtidas na etapa anterior;
- Analisar as capacidades discriminativas das estratégias em comparação com os modelos disponíveis na literatura especializada;
- Definir as principais associações e limites observados no contexto aqui explorado.

Considerando o previsto em Edital PROPe Unesp Nº 08/2024 - PIBIC, esta proposta tem aderência plena com as áreas Prioritárias do Ministério da Ciência, Tecnologia, Inovações e Comunicações (estabelecidas na Portaria MCTIC nº 1.122/2020, com texto alterado pela Portaria MCTIC nº 1.329/2020), especificamente com as Áreas Tecnologias Estratégicas (Cibernética) e Tecnologias Habilitadoras (Inteligência Artificial). No que tange a lista Objetivos do Desenvolvimento Sustentável (ODS), essa pesquisa apresenta aderência direta com a ODS 9 (indústria, inovação e infraestrutura).

3 Metodologia

A pesquisa proposta será desenvolvida em etapas, descritas a seguir:

3.1 Etapa 1 - Revisão Bibliográfica

Esta etapa permite o levantamento bibliográfico necessário para manter o projeto atualizado, proporcionar uma fundamentação teórica sólida e subsidiar a exploração proposta.

3.2 Etapa 2 - Exploração de *BERT*

Esta etapa é direcionada para explorar a arquitetura BERT em razão dos resultados conquistados na área de processamento de linguagem natural, especialmente a partir de variações voltadas para a cybersegurança [**?**]. O modelo que será analisado e investigado foi descrito por Seyyar et al. [**?**], em razão dos relevantes resultados conquistados para a detecção de *CyberAtaques* através de pedidos HTTP/HTTPS. O modelo foi fundamentado em BERT para a criação dos vetores de palavras e MLP para a classificação. Buscando viabilizar a aplicação do modelo BERT em questão, serão examinadas amostras de ataques dos *distributed denial of service* (DDoS), *man in the middle* (MitM), SQL injection e *cross site scripting* (XSS) [**?**]. A ideia é analisar estes contextos a fim de verificar a viabilidade do uso de abordagens fundamentadas em sentenças regulares e mecanismos explorados via BERT. Ao comparar a entrada com as amostras maliciosas ou não presentes nos *datasets* utilizados, a etapa de MLP retornará a classificação do *input* fornecido, indicando assim, se alguma ameaça está ou não presente na amostra, e qual o tipo de ataque. Outras estratégias podem ser exploradas nesta etapa a fim de aprimorar e/ou atualizar a proposta em observação as possíveis tendências observadas na literatura especializada.

3.3 Etapa 3 - Exploração de modelo *ViT*

Modelos ViT representam uma variação de *transformers*, voltada para o processamento de imagens. Nesse caso, a composição de *feature vectors* é definida dividindo a imagem examinada em *patches*. Existem diversas formas de realizar essa divisão, porém a mais usual é o formato 16x16 pixels. Após os *patches* serem devidamente distribuídos, uma transformação linear é aplicada, semelhantemente ao BERT, gerando assim os respectivos *embeddings* de posições. Em seguida, os embeddings são analisados por meio de várias camadas, formando uma estrutura de *multi-head attention*. Com isso, cada *patch* de imagem é comparado aos outros, de diferentes formas, atribuindo, portanto, múltiplos valores aos *feature vectors*, tornando o processo de classificação mais preciso [**?**].

No contexto de aplicação desta proposta, as amostras de dados (ataque ou quebra de segurança de dados) encontram-se no formato de texto. Por isso, é necessário realizar a conversão dos textos para imagem para viabilizar o uso de ViT. Neste trabalho, o modelo que será investigado foi nomeado como *Butterfly Vision Transformer*(B_ViT) [**?**]. É válido ressaltar que o B_ViT utiliza uma arquitetura ViT pré-treinada, e o que o difere dos demais é o seu mecanismo de *attention*, que também foi modificado. Além de utilizar como entrada códigos executáveis convertidos em imagens em níveis de cinza, B_ViT opera com dois tipos de *attention*, sendo uma delas a *local attention*, e a outra *global attention*. *local attention* será aplicada em pequenas partes da imagem, os patches, com a intenção de detectar detalhes em uma área restrita, resultando em melhor exploração da imagem dada. Isso permite quantificar informações maliciosas ocultas ou mascaradas. Já a parte de *global attention* será aplicada na imagem completa. Ao aplicar *global attention*, será verificado se o modelo consegue entender a relevância e a abrangência de cada patch, relacionando-os a fim de obter uma interpretação mais ampla sobre um malware. As vantagens em utilizar o B_ViT se baseiam em capturar tanto detalhes finos quanto padrões amplos, resultando em uma análise mais precisa e completa para a classificação de imagens de exemplos de *malwares*.

Para viabilizar esta etapa, é necessário aplicar técnicas para transformação de texto em imagens (*reshaping*), que se baseiam em transformações lineares aplicadas sobre a representação dos textos geradas pelo modelo. Algumas técnicas que serão exploradas neste trabalho visam permitir o uso do B_ViT para classificar e reconhecer amostras maliciosas presentes em diferentes *datasets*, tais como *sequential reshape*, *recurrence plot*, *gramian angular field* (GAF) [**?**], e imagens em níveis de cinza. As imagens geradas após a aplicação do processo de *reshaping* serão passadas para o modelo B_ViT como entrada, possibilitando assim, outra perspectiva quanto à classificação de comportamentos intrusivos presentes nos códigos, que antes eram representados e analisados apenas como texto.

4 Etapa 4 - Contexto de Aplicação: datasets

Todos os tipos de dados utilizados no processo de treinamento e aprendizado de um modelo encontram-se armazenados e organizados em um dataset, separados por classes. No contexto do projeto apresentado, os datasets utilizados pelos modelos pré-treinados escolhidos estão concentrados na área da *CyberSegurança*, tais como amostras de ataques dos *distributed denial of service* (DDoS), *man in the middle* (MitM), SQL injection e *cross site scripting* (XSS) [**?**]. Logos, a ideia é verificar a viabilidade das associações previstas neste trabalho nos seguintes datasets: CSIC 2010 [**?**], FWAF [**?**] e *httpParams* [**?**]. Ao implementar a aplicação BERT, após o registro dos resultados obtidos, os dados de texto que representam os variados ataques cibernéticos passarão pela conversão de texto para imagem. Logo, as técnicas de *reshaping* serão aplicadas em todas as amostras contidas nos datasets, gerando, portanto, um novo banco de imagens, que servirá como base de estudos para o modelo B_ViT em questão.

4.1 Etapa 5 - Análises e Extração de Conhecimento

O processo de análise e avaliação de desempenho ocorrerá por meio de métricas comumente exploradas na área de aprendizado de máquina, tais como acurácia, *medida-F*, *recall* e outras

[**?**]. A validação da melhor combinação ocorrerá por meio de comparações dos desempenhos com: fornecidos por trabalhos correlatos; e, obtidas via as diferentes arquiteturas indicadas nesta proposta.

5 Plano de Trabalho e Cronograma de Execução

Este projeto é parte de outros projetos que estão em pleno desenvolvimento pelo orientador. O plano de trabalho consiste em realizar as etapas descritas na seção 3: Etapa 1 - Revisão Bibliográfica; Etapa 2 - Exploração de *BERT*; Etapa 3 - Exploração de modelo *ViT*; Etapa 4 - Contexto de Aplicação: datasets; Etapa 5 - Análises e Extração de Conhecimento. O cronograma de execução proposto para 12 meses está na Tabela 1.

Tabela 1: Proposta de Cronograma para o atendimento das etapas previstas no escopo do projeto, considerando o período de 12 meses.

Etapas	Ano 1: 2024-2025											
	1	2	3	4	5	6	7	8	9	10	11	12
1	x		x		x		x		x			
2	x	x	x									
3				x	x							
4					x	x						
5							x	x	x	x		
Escrita: Processos e Resultados		x		x		x		x		x	x	x

6 Resultados Esperados

Espera-se que a execução deste projeto forneça resultados relevantes, tais como: viabilidade do uso de um modelo ViT para classificar e reconhecer padrões na área de *CyberAtaques*, tais como por meio de requisições HTTP/HTTPS; conhecimento sobre as principais associações (*reshaping* com ViT) e limites correspondentes no contexto aqui explorado; indicação da capacidade discriminativa mais relevante entre BERT e B_ViT para os conjuntos de dados considerados nos experimentos práticos destes modelos, contribuindo para a compreensão e avanço no campo de análise de técnicas direcionadas para *CyberAtaques*.