Comparação entre Random Forest e LSTM na Detecção de Ataques DDoS

Comparison between Random Forest and LSTM for DDoS Attack Detection

Guilherme Henrique Soeiro Fontes[1], Luiz Fernando Carvalho[2]

RESUMO

Este artigo apresenta uma comparação entre os algoritmos Random Forest e LSTM (*Long Short-Term Memory*) para detecção de ataques DDoS (*Distributed Denial of Service*). Os modelos foram desenvolvidos no ambiente do Google Colab e os códigos-fonte estão disponibilizados para referência. Surpreendentemente, a Random Forest demonstrou eficiência comparável à LSTM na detecção de ataques DDoS, apesar de sua simplicidade em relação aos requisitos computacionais. A LSTM, conhecida por sua capacidade de modelar sequências temporais complexas, mostrou alto custo computacional em termos de tempo de execução e consumo de memória RAM. Esse achado enfatiza a importância de considerar não apenas a precisão dos modelos, mas também seus recursos computacionais necessários ao escolher uma abordagem de detecção de DDoS. A simplicidade e eficiência da Random Forest podem torná-la uma solução prática e acessível, especialmente em ambientes com recursos limitados.
PALAVRAS-CHAVE: DDoS; Long Short-Term Memory; Random Forest.

ABSTRACT

This work presents a comparison between the Random Forest and LSTM (Long Short-Term Memory) algorithms for detecting Distributed Denial of Service (DDoS) attacks. The models were developed in the Google Colab environment, and the source codes are made available for reference. Surprisingly, Random Forest demonstrated efficiency comparable to LSTM in detecting DDoS attacks, despite its simplicity in terms of computational requirements. LSTM, known for its ability to model complex temporal sequences, exhibited high computational cost in terms of execution time and RAM consumption. This finding emphasizes the importance of considering not only the accuracy of the models but also their computational resources when choosing a DDoS detection approach. The simplicity and efficiency of Random Forest can make it a practical and accessible solution, especially in resource-constrained environments.
KEYWORDS: DDoS; Long Short-Term Memory; Random Forest.

INTRODUÇÃO

A crescente complexidade e sofisticação dos ataques cibernéticos, notadamente os Distribuídos de Negação de Serviço (DDoS), representam um desafio global significativo para a segurança de sistemas e redes digitais. A detecção e a mitigação eficazes desses ataques são vitais para assegurar a disponibilidade e a integridade de serviços online. Nesse cenário, técnicas de aprendizado de máquina têm ganhado destaque como uma abordagem promissora na identificação e resolução de ataques DDoS, dada a complexidade inerente à natureza distribuída dessas ações (RANI et al., 2022). Esses ataques têm como alvo principal o aumento do tráfego de rede, seja por

[1] Bolsista do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Universidade Tecnológica Federal do Paraná, Apucarana, Paraná, Brasil. E-mail: guilhermehenriquefontes@alunos.utfpr.edu.br. ID Lattes: 9914658536924375.
[2] Docente no curso de Engenharia de Computação. Universidade Tecnológica Federal do Paraná Apucarana, Paraná, Brasil. E-mail:luizfcarvalho@utfpr.edu.br. ID Lattes:0665079937865380.

meio do aumento do volume de pacotes ou da geração de um grande volume de mensagens de controle.

Este trabalho tem como objetivo central a comparação de dois algoritmos de aprendizado de máquina, *Random Forest* e *Long Short-Term Memory* (LSTM), quanto ao seu desempenho na detecção de ataques DDoS. A Random Forest é uma técnica de ensemble que se destaca por sua capacidade de lidar com uma variedade de tipos de dados e é conhecida por sua eficácia na classificação. As Random Forest apresentam uma certa eficiência na classificação de ataques DDoS (BOUKE et al., 2023). Por outro lado, as redes LSTM são amplamente reconhecidas por sua habilidade em modelar sequências temporais complexas, o que é relevante na detecção de padrões de ataques DDoS em fluxos de tráfego de rede.

A investigação comparativa desses dois algoritmos é de grande relevância, uma vez que pode fornecer uma compreensão sobre suas capacidades e limitações no contexto da segurança cibernética. Além disso, contribuirá para o desenvolvimento de soluções mais eficazes e precisas na detecção de ataques DDoS, melhorando a resiliência de sistemas e serviços digitais.

Neste trabalho, foram coletados dados de tráfego de rede representativos de cenários de ataques DDoS reais. Esses dados serão utilizados para treinar e testar os modelos baseados em Random Forest e LSTM. A métrica de desempenho será avaliada com base na precisão, revocação, F1-score e outros indicadores relevantes.

MATERIAIS E MÉTODOS

A pesquisa foi conduzida em um ambiente composto por um único computador com acesso à internet, utilizando o ambiente virtual Google Colab para a execução dos códigos em Python. O conjunto de dados utilizado para esta pesquisa, "CIC DDoS 2019", foi obtido da Universidade de New Brunswick, no Canadá, disponível em https://www.unb.ca/cic/datasets/ddos-2019.html.

É importante ressaltar que o dataset usado nesta pesquisa foi construído por meio da coleta de dados de uma rede real, na qual foram realizados 10 diferentes tipos de ataques DDoS. A coleta de dados ocorreu ao longo de dois dias, sendo que os dados fornecidos foram pré-processados para agrupar o tráfego da rede segundo a segundo. O dataset resultante contém mais de 16 mil amostras/linhas, sendo que cada amostra é composta por 87 atributos. Desses atributos, 86 representam características do tráfego de rede, como quantidade de pacotes, bytes e outras informações de controle, enquanto o último atributo é um rótulo que indica se ocorreu ou não um ataque durante aquele segundo amostrado. Antes de sua utilização na pesquisa, os dados passaram por um processo de pré-processamento rigoroso, que incluiu a verificação de dados corrompidos ou ausentes.

Após a etapa de pré-processamento, realizamos uma análise de correlação entre os 86 atributos e o resultado dos ataques, mantendo apenas aqueles com uma correlação maior ou igual ao valor absoluto de 0,5. Isso resultou na seleção de 14 atributos considerados relevantes para a pesquisa. A seleção desses atributos é para diminuir o custo computacional da rede neural, já que na LSTM, trabalhar com todos eles seria inviável.

Para avaliar o desempenho da detecção de ataques DDoS, foram testados nove modelos LSTM (*Long Short-Term Memory*) com diferentes configurações. Essas configurações envolveram variações no tamanho do histórico na série temporal, com experimentos conduzidos em intervalos de 30, 60 e 90 segundos. Além disso, foram exploradas diversas taxas de esquecimento em cada camada dos modelos, com valores testados incluindo 0,2, 0,3 e 0,4. Esse enfoque no uso de LSTM com série temporal é uma estratégia que proporciona desempenho do sistema com tempos de resposta rápidos e alta precisão no reconhecimento (CHU et al., 2021).

Também foram utilizados dois modelos Random Forest, um com todos os atributos e outro apenas com os mesmos 14 atributos selecionados para as LSTM. Para a avaliação do desempenho, foi analisada a quantidade de resultados que foram classificados como positivos e negativos e quantos foram classificados corretamente, Dessa forma também foi possível calcular métricas como F1-Score, Acurácia, Precisão e Revocação. Além disso, elaboramos tabelas com os resultados obtidos por cada modelo.

É importante ressaltar que os parâmetros, como funções de ativação, otimizadores, número de camadas e neurônios da LSTM, foram escolhidos após uma série de testes, visando alcançar um desempenho satisfatório e um custo computacional gerenciável. Embora reconheçamos que essa configuração não seja necessariamente a ideal em todas as circunstâncias, nosso foco principal estava na análise abrangente das abordagens de aprendizado de máquina na detecção de ataques DDoS, considerando eficácia e custos computacionais.

Este processo metodológico permitiu uma análise aprofundada das abordagens de aprendizado de máquina na detecção de ataques DDoS, considerando eficácia e custos computacionais, e enriqueceu nosso entendimento das melhores práticas nessa área.

RESULTADOS E DISCUSSÕES

Após a montagem dos modelos e a realização dos testes, como detalhado na seção anterior, procedemos à análise dos dados classificados pelo algoritmo em relação a sua verdadeira classificação. Para facilitar o entendimento, a classificação do tráfego foi dividida em quatro categorias: Verdadeiros Positivos (acertos de casos positivos), Verdadeiros Negativos (acertos de casos negativos), Falsos Positivos (erros de classificação positiva) e Falsos Negativos (erros de classificação negativa).

Registramos todos esses valores, que estão disponíveis na Tabela 1. Na tabela, "DropOut" representa a taxa de esquecimento definida para cada camada da LSTM, e

"Qtde. dados" indica a quantidade de dados de histórico utilizada para a montagem do conjunto de dados.

Tabela 1 – Tabela de Resultados

Modelo	Qtde. Dados	DropOut	VP	FP	VN	FN
LSTM	30	2	$1{,}8 \times 10^3$	$5{,}9 \times 10^2$	$4{,}3 \times 10^3$	5×10^2
LSTM	60	2	$2{,}2 \times 10^3$	$5{,}2 \times 10^2$	$4{,}3 \times 10^3$	$1{,}1 \times 10^2$
LSTM	90	2	$1{,}8 \times 10^3$	$2{,}7 \times 10^2$	$4{,}6 \times 10^3$	$5{,}2 \times 10^2$
LSTM	30	3	$2{,}2 \times 10^3$	$2{,}2 \times 10^3$	$2{,}6 \times 10^3$	$1{,}3 \times 10^2$
LSTM	60	3	$2{,}2 \times 10^3$	$6{,}3 \times 10^2$	$4{,}2 \times 10^3$	$1{,}3 \times 10^2$
LSTM	90	3	$2{,}2 \times 10^3$	$6{,}2 \times 10^2$	$4{,}2 \times 10^3$	$1{,}3 \times 10^2$
LSTM	30	4	$2{,}2 \times 10^3$	$9{,}7 \times 10^2$	$3{,}9 \times 10^3$	$1{,}2 \times 10^2$
LSTM	60	4	$1{,}8 \times 10^3$	$1{,}6 \times 10^2$	$4{,}7 \times 10^3$	$5{,}1 \times 10^2$
LSTM	90	4	$1{,}8 \times 10^3$	$5{,}9 \times 10^2$	$4{,}3 \times 10^3$	5×10^2
R. F. (14)	–	–	$2{,}2 \times 10^3$	$5{,}4 \times 10^2$	$4{,}3 \times 10^3$	99
R. F. (ALL)	–	–	$2{,}2 \times 10^3$	$6{,}5 \times 10^2$	$4{,}2 \times 10^3$	59

Fonte: Elaborado pelos autores

É evidente que todos os modelos demonstraram um desempenho satisfatório, com eficiência acima de 60%. Em todos os casos, o número de classificações corretas foi três vezes maior do que as classificações incorretas. Com base nessas informações, calculamos a Acurácia, Precisão, Revocação e F1-Score para cada modelo. Esses valores estão organizados na Tabela 2.

Tabela 2 – Resultados das Métricas para Avaliação

Modelo	Qtde. Dados	DropOut	Acurácia	Precisão	Revocação	F1-Measure
LSTM	30	2	0,8475	0,7512	0,7795	0,7651
LSTM	60	2	0,9113	0,8057	0,9526	0,8730
LSTM	90	2	0,8882	0,8672	0,7699	0,8156
LSTM	30	3	0,6705	0,4911	0,9443	0,6461
LSTM	60	3	0,8938	0,7735	0,9447	0,8506

LSTM	90	3	0,8953	0,7778	0,9439	0,8529
LSTM	30	4	0,8606	0,7085	0,9557	0,8138
LSTM	60	4	0,8460	0,6891	0,9452	0,7971
LSTM	90	4	0,9053	0,9162	0,7765	0,8406
R. F. (14)	–	–	0,9108	0,8028	0,9570	0,8731
R. F. (ALL)	–	–	0,9015	0,7758	0,9744	0,8638

Fonte: Elaborado pelos autores

Contrariando as expectativas iniciais, os resultados obtidos não revelaram um padrão claro nas métricas de desempenho das LSTM. Inicialmente, supúnhamos que ao aumentar a taxa de DropOut entre as camadas, a eficiência diminuiria, e ao adicionar mais segundos como histórico na montagem do conjunto de dados, a eficiência aumentaria de forma consistente. Entretanto, ao examinarmos as métricas apresentadas na Tabela 2, fica evidente que os resultados variaram sem demonstrar um padrão discernível que justificasse essas suposições.

Um ponto de interesse é a semelhança de desempenho entre ambas as Random Forests, com uma leve vantagem para aquelas com menos parâmetros. Isso sugere que nossa análise no pré-processamento dos dados foi eficaz na seleção de atributos relevantes. Surpreendentemente, as Random Forests superaram as LSTM em termos de desempenho, destacando que a quantidade de histórico usada nas LSTM não parece ter uma influência significativa sobre os resultados finais. A variação na taxa de perda (*Drop Out*) e o aumento no número de segundos de histórico não produziram os efeitos esperados.

CONCLUSÃO

Com base nas análises e resultados obtidos, é possível concluir que a Random Forest com os 14 atributos se destaca como a alternativa de melhor desempenho entre todos os modelos testados. Suas métricas de avaliação apresentaram resultados consistentes e superiores, mesmo quando comparadas à versão que utiliza todos os atributos durante o treinamento. Além disso, a Random Forest demonstrou uma vantagem significativa em termos de tempo de treinamento, com um período de treinamento que não excedeu 15 segundos, em contraste com os modelos LSTM, que demandaram mais de 30 minutos para treinamento.

É importante ressaltar que todas essas conclusões são baseadas nos resultados analisados neste estudo específico, e podem ser sujeitas a variações em trabalhos futuros. Variações podem ocorrer à medida que se exploram diferentes combinações de atributos nos algoritmos de Machine Learning, estruturas alternativas para as LSTM e a

inclusão de outros algoritmos que utilizam dados históricos para classificação, como o ARIMA.

No entanto, os resultados aqui apresentados fornecem uma base sólida e promissora para a detecção de ataques DDoS usando a Random Forest com seleção criteriosa de atributos. Essa abordagem se destaca não apenas pelo desempenho sólido, mas também pela eficiência computacional, o que a torna uma candidata viável para aplicações práticas de detecção de ataques DDoS. O progresso contínuo nessa área exigirá investigações mais aprofundadas e experimentações adicionais, mas os resultados deste estudo abrem caminho para futuras melhorias e avanços na detecção de ameaças cibernéticas.

Disponibilidade do Código

É possível acessar o código e os recursos relacionados por meio do seguinte link: https://github.com/KHKGuilherme/LongShortTermMemory/tree/main

CONFLITO DE INTERESSE

Não há conflito de interesse.

AGRADECIMENTOS

Gostaria de expressar meus profundos agradecimentos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela bolsa de Iniciação Científica que tornou possível este projeto. Sou imensamente grato ao meu orientador, Luiz Fernando Carvalho, por suas contribuições essenciais ao sucesso do projeto e meu desenvolvimento como pesquisador. Essa oportunidade não apenas aliviou as preocupações financeiras, mas também me permitiu crescer como pesquisador, explorando áreas previamente desconhecidas e, de maneira fundamental, incentivando e estimulando meu entusiasmo pela pesquisa acadêmica.

REFERÊNCIAS

CHU, H.-C.; YAN, C.-Y. **DDoS Attack Detection with Packet Continuity Based on LSTM Model**. In: IEEE 3rd Eurasia Conference on IOT, Communication and Engineering (ECICE), 2021, Yunlin, Taiwan. Proceedings. Yunlin, Taiwan, 2021. p. 44-47. DOI: 10.1109/ECICE52819.2021.9645650.

BOUKE, Mohamed Aly et al. **An intelligent DDoS attack detection tree-based model using Gini index feature selection method**. Microprocessors and Microsystems, v. 98, 2023, p. 104823. ISSN 0141-9331.

RANI, S.V. Jansi et al. **Detection of DDoS attacks in D2D communications using machine learning approach**. Computer Communications, v. 198, 2023, p. 32-51. ISSN 0140-3664.