Bibliotecas de Software Livre para Detecção de Esteganografia em Imagens Digitais

Érico Meger[1], **Eros Henrique Lunardon Andrade**[1], **Guilherme Werneck de Oliveira**[1]

[1]Campus Pinhais – Instituto Federal do Paraná (IFPR) Pinhais - PR - Brasil

Abstract. *This meta-paper describes the style to be used in articles and short papers for SBC conferences. For papers in English, you should add just an abstract while for the papers in Portuguese, we also ask for an abstract in Portuguese ("resumo"). In both cases, abstracts should not have more than 10 lines and must be in the first page of the paper.*

Resumo. *Este meta-artigo descreve o estilo a ser usado na confecção de artigos e resumos de artigos para publicação nos anais das conferências organizadas pela SBC. É solicitada a escrita de resumo e abstract apenas para os artigos escritos em português. Artigos em inglês deverão apresentar apenas abstract. Nos dois casos, o autor deve tomar cuidado para que o resumo (e o abstract) não ultrapassem 10 linhas cada, sendo que ambos devem estar na primeira página do artigo.*

1. Introdução

O movimento do software livre se estabelece como um paradigma essencial para promover transparência, colaboração e inovação no cenário tecnológico contemporâneo. Segundo a Free Software Foundation, software livre é definido pela sua capacidade de respeitar as liberdades e o controle dos usuários sobre o software: a liberdade de executar o programa para qualquer propósito, de estudá-lo e modificá-lo (acesso ao código-fonte é pré-requisito), de redistribuir cópias e de distribuir versões modificadas para a comunidade, conhecidas como as quatro liberdades essenciais [Foundation 2024].

Ao assegurar essas liberdades, o software livre não apenas fortalece a confiança nas soluções digitais, por permitir auditoria e aprendizado mútuo, mas também fomenta ambientes colaborativos dinâmicos, onde ferramentas podem ser aprimoradas coletivamente. Essa filosofia de abertura e colaboração se manifesta também no campo da inteligência artificial, por meio de bibliotecas como PyTorch, TensorFlow e scikit-learn. Essas ferramentas de código aberto democratizam o acesso a algoritmos de aprendizado de máquina, permitindo reprodutibilidade científica, auditoria de modelos e desenvolvimento colaborativo de soluções inovadoras [Team 2025a, Team 2025b].

No contexto da esteganografia, a disponibilidade dessas bibliotecas open source oferece grandes oportunidades para o avanço da área. A análise de imagens digitais, por exemplo, pode se beneficiar de recursos de detecção de padrões e classificação automática fornecidos por essas ferramentas, auxiliando tanto no desenvolvimento de técnicas esteganográficas mais robustas quanto na criação de métodos de detecção mais eficazes. Assim, a intersecção entre software livre, inteligência artificial e esteganografia evidencia como a filosofia do código aberto não só fortalece a confiança técnica, mas também amplia as possibilidades de pesquisa e aplicação prática neste campo.

A esteganografia pode ser compreendida como uma técnica utilizada para esconder informações em meios aparentemente comuns, de forma que um observador externo não consiga identificar a presença de dados ocultos [Fridrich 2010].

Essa área de estudo, portanto, não se limita apenas ao ato de esconder informações, mas constitui um campo de estudo mais amplo que abrange técnicas, algoritmos e aplicações destinadas a garantir a confidencialidade e a discrição da comunicação. Em contraste com a criptografia, que protege o conteúdo das mensagens mas não oculta sua existência, a esteganografia busca mascarar o próprio ato de comunicação [Fridrich 2010]. Essa característica a torna uma área estratégica tanto para aplicações legítimas, como autenticação de documentos e proteção da privacidade, quanto para usos maliciosos. Tal dualidade evidencia que a esteganografia deve ser compreendida não apenas sob uma perspectiva técnica, mas também dentro de um contexto social e político mais amplo.

Nesse sentido, ao longo da história, e de forma ainda mais acentuada no cenário contemporâneo, observa-se o fortalecimento de mecanismos de vigilância e controle sobre a comunicação digital. Na Europa, por exemplo, esse movimento se materializa tanto em iniciativas de remoção massiva de conteúdos, com mais de 41 milhões de postagens bloqueadas apenas no primeiro semestre de 2025 [Poder360 2025], quanto em pressões políticas para enfraquecer a segurança criptográfica, como a exigência de um backdoor no iCloud, que levou a Apple a retirar a opção de criptografia de ponta a ponta de seus serviços no Reino Unido [Guardian 2025]. Embora tais medidas sejam frequentemente justificadas em nome da segurança pública, a ausência de transparência sobre os critérios de censura e o impacto direto na privacidade digital levantam sérias preocupações. Nesse contexto, a esteganografia age como uma alternativa tecnológica de resistência, capaz de proporcionar meios de comunicação discretos e seguros, reforçando sua relevância sociopolítica e justificando o aprofundamento de seu estudo.

1.1. Objetivo

Explorar o uso de bibliotecas de software livre no desenvolvimento de modelos de inteligência artificial para a detecção de esteganografia em imagens digitais.

2. Revisão bibliográfica

Essa seção revisará os principais trabalhos relacionados, destacando contribuições, métodos e limitações que fundamentam o desenvolvimento desta pesquisa.

O trabalho *"An Ensemble Model using CNNs on Different Domains for ALASKA2 Image Steganalysis"* de Chubachi [Chubachi 2020] surge da constatação de que muitos detectores de esteganografia baseados em aprendizado profundo não generalizam bem em cenários reais devido ao uso de conjuntos de dados simplificados. A competição ALASKA2 ofereceu um ambiente mais realista, com imagens JPEG coloridas de diferentes origens e processos, estimulando soluções mais aplicáveis. Nesse contexto, o objetivo do autor foi desenvolver um modelo de detecção baseado em um ensemble de redes convolucionais que combinasse informações tanto do domínio espacial (RGB, YUV e Lab) quanto do domínio da frequência (coeficientes DCT).

A metodologia proposta envolveu CNNs construídas sobre arquiteturas EfficientNet, com ajustes para lidar com as especificidades de cada domínio. No caso dos coe-

ficientes DCT, foram aplicadas codificações one-hot, recortes de valores e convoluções dilatadas para capturar padrões sutis. Para integrar os modelos, além da simples média de previsões, foi desenvolvido um perceptron multicamada capaz de combinar os mapas de características. Também se utilizaram técnicas auxiliares, como pseudo-rotulagem e stacking com LightGBM. Em experimentos conduzidos com 300 mil imagens, o uso combinado dos modelos trouxe ganhos consistentes, resultando em uma performance de AUC ponderado próxima de 0,94 e garantindo a terceira colocação na competição.

O estudo apresenta como pontos fortes a inovação de combinar diferentes domínios e a validação em um cenário competitivo e realista. Contudo, o alto custo computacional e a limitação de testar apenas algoritmos de esteganografia já conhecidos restringem sua aplicabilidade prática. Em contraste, a proposta desenvolvida neste trabalho busca explorar o uso de bibliotecas de software livre para a construção de modelos de inteligência artificial em esteganálise.

O trabalho *ImageNet Pre-trained CNNs for JPEG Steganalysis* de Yousfi et al. apresenta uma investigação sobre o uso de redes neurais convolucionais (CNNs) pré-treinadas para a esteganálise de imagens JPEG. A principal motivação para o estudo surgiu a partir da competição de esteganálise ALASKA II, onde foi observado que os participantes com melhor desempenho utilizavam modelos de visão computacional de uso geral, como EfficientNet e ResNet, em vez de arquiteturas especializadas e treinadas do zero para a tarefa [Yousfi et al. 2021]. Essa nova abordagem, baseada em aprendizagem por transferência (transfer learning), representava uma mudança de paradigma em relação a modelos consolidados, como a SRNet, que eram projetados especificamente para esteganálise [Yousfi et al. 2021]. Diante desse cenário, o principal objetivo dos autores foi investigar e demonstrar formalmente a eficácia e a superioridade desses modelos pré-treinados no ImageNet para a detecção de esteganografia em imagens JPEG, comparando seu desempenho com as abordagens tradicionais [Yousfi et al. 2021].

A metodologia utilizada centrou-se em refinar (fine-tuning) diversas arquiteturas pré-treinadas, como EfficientNet, MixNet e ResNet, no conjunto de dados da ALASKA II [Yousfi et al. 2021]. Os autores também conduziram experimentos para avaliar o impacto de decisões arquitetônicas, como a remoção de camadas de pooling ou stride no início da rede, confirmando que a manutenção da resolução original nas primeiras camadas é crucial para a performance em esteganálise [Yousfi et al. 2021]. O experimento principal foi conduzido no dataset da ALASKA II, que continha imagens comprimidas com fatores de qualidade 75, 90 e 95 e com mensagens ocultas pelos algoritmos J-UNIWARD, J-MiPOD e UERD [Yousfi et al. 2021]. Os resultados demonstraram que os modelos pré-treinados ofereceram um desempenho de detecção significativamente superior ao da SRNet em todas as configurações testadas [Yousfi et al. 2021].

Os pontos fortes dessa abordagem, destacados no artigo, são a acurácia superior, a maior eficiência de dados e uma velocidade de treinamento ordens de magnitude mais rápida em comparação com o treinamento de um modelo do zero [Yousfi et al. 2021]. Como limitação, os próprios autores apontam que o estudo foi amplamente focado no ambiente da ALASKA II e que o ganho de desempenho obtido com o pré-treinamento tende a diminuir à medida que o volume de dados para treinamento na tarefa final aumenta [Yousfi et al. 2021].

O estudo de *Ensemble of CNNs for Steganalysis: An Empirical Study* de Xu et al. é motivado pela observação de que, embora as Redes Neurais Convolucionais (CNNs) estivessem ganhando popularidade em esteganálise, a maioria das pesquisas se concentrava no design de um único modelo de CNN. No entanto, no campo mais amplo da visão computacional e do aprendizado de máquina, as melhores performances são consistentemente alcançadas por meio de ensembles, ou seja, a combinação de múltiplos modelos [Xu et al. 2016]. Os autores perceberam uma lacuna na literatura de esteganálise, que ainda não havia explorado a fundo estratégias de ensemble mais sofisticadas do que a simples média das previsões. O principal objetivo do trabalho foi, portanto, conduzir um estudo empírico para avaliar o desempenho de diferentes estratégias de combinação de CNNs para a tarefa de esteganálise, buscando ir além da média de modelos e testar o uso de classificadores de segundo nível treinados sobre as saídas e representações internas das redes [Xu et al. 2016].

A metodologia proposta envolveu, primeiramente, o treinamento de um conjunto de 16 CNNs individuais, que serviram como ”aprendizes de base”, cada uma treinada sobre um subconjunto aleatório do dataset de treinamento [Xu et al. 2016]. A partir disso, os autores testaram e compararam três abordagens de ensemble: a média simples das probabilidades de saída, a criação de novos vetores de características a partir dessas probabilidades para treinar um classificador de segundo nível, e uma terceira técnica mais inovadora que consistia em extrair as representações de características das camadas intermediárias de cada CNN (antes da camada final de classificação), concatená-las e usá-las para treinar o classificador de segundo nível [Xu et al. 2016]. O experimento foi realizado no dataset BOSSbase v1.01 para detectar a esteganografia do algoritmo S-UNIWARD com uma taxa de inserção de 0.4 bpp, utilizando duas arquiteturas de CNN de tamanhos diferentes para avaliar o impacto da capacidade do modelo [Xu et al. 2016].

O trabalho apresenta como pontos fortes a demonstração sistemática de que o uso de um classificador de segundo nível sobre as saídas das CNNs melhora consistentemente o desempenho em relação à média de modelos, e, principalmente, a descoberta de que o uso das representações de características intermediárias resulta na melhor performance, indicando que os classificadores mais robustos podem extrair padrões mais discriminativos do que as camadas de classificação simples das CNNs base [Xu et al. 2016]. As limitações do estudo, no entanto, incluem a sua validação em apenas um dataset, contra um único algoritmo esteganográfico e com uma única taxa de inserção, além do alto custo computacional da abordagem [Xu et al. 2016].

O artigo *An Intriguing Struggle of CNNs in JPEG Steganalysis and the OneHot Solution* de Yousfi e Fridrich [Yousfi and Fridrich 2020] parte de uma motivação muito clara: a descoberta de cenários específicos onde as modernas Redes Neurais Convolucionais (CNNs), como a SRNet, apresentavam um desempenho surpreendentemente inferior ao de métodos mais antigos baseados em extração de características, como o JPEG Rich Model (JRM). Essa falha era particularmente evidente na detecção do algoritmo nsF5 e do J-UNIWARD em certos tipos de imagem JPEG, e a análise revelou que o sucesso do JRM nesses casos se devia à sua capacidade de computar estatísticas simples dos coeficientes DCT, como histogramas de coocorrência, algo que as CNNs convencionais, que operam na imagem descomprimida, não conseguiam ”enxergar”. O objetivo principal do trabalho foi, portanto, diagnosticar essa deficiência e propor uma nova ar-

quitetura de CNN, chamada ”OneHot”, que fosse capaz de aprender diretamente a partir dos coeficientes DCT e, adicionalmente, desenvolver uma metodologia para integrar essa nova rede a arquiteturas já existentes, criando um detector mais universal e robusto.

Para alcançar esse objetivo, a metodologia se baseia em duas inovações principais. A primeira é a própria rede ”OneHot CNN”, que transforma sua entrada através de uma camada de ”codificação one-hot com corte” (clipped one-hot encoding). Nessa etapa, os valores absolutos dos coeficientes DCT são transformados em um volume binário que facilita o aprendizado de ocorrências e coocorrências pelas camadas convolucionais subsequentes, que utilizam uma combinação de convoluções padrão e dilatadas para capturar relações estatísticas. A segunda inovação é a arquitetura de ramo duplo ”OneHot+SRNet”, que mescla a nova rede OneHot com uma CNN convencional (SRNet). Cada rede opera em um ramo paralelo, e as representações de características de ambos são concatenadas antes de uma camada de classificação final, permitindo que o modelo combinado seja treinado de ponta a ponta. O experimento principal consistiu em testar o desempenho dessas novas arquiteturas nos cenários problemáticos (nsF5 e J-UNIWARD) em datasets como BOSSbase e BOWS2, onde a OneHot CNN sozinha superou o JRM, e a OneHot+SRNet melhorou substancialmente a SRNet, sem prejudicar seu desempenho em casos onde já era eficaz.

O trabalho se destaca por identificar com precisão uma falha em modelos estado da arte e propor uma solução elegante e eficaz, a codificação one-hot, que permite a uma CNN aprender estatísticas de alta ordem de forma flexível. A arquitetura de ramo duplo é outro ponto forte, pois oferece uma maneira prática de criar um detector mais completo e robusto. Uma limitação implícita é que a rede OneHot é altamente especializada para esses casos de falha, o que justifica sua fusão com uma rede mais geral como a SRNet, em vez de ser usada isoladamente para todos os cenários.

O artigo ”Comprehensive survey on image steganalysis using deep learning” de De La Croix et al. [La Croix et al. 2024] surge como uma resposta à evolução e complexidade do campo da esteganálise. A principal motivação dos autores reside na observação de que as abordagens tradicionais, baseadas em aprendizado de máquina (ML), se mostraram ineficazes contra os modernos e prevalentes algoritmos de esteganografia [La Croix et al. 2024]. Métodos clássicos como Support Vector Machines (SVM) e Ensemble Classifiers (EC) dependem de um processo árduo e manual de extração de características (conhecidos como rich models, como o SRM), o que não só consome tempo, mas também sofre com a ”maldição da dimensionalidade”, onde o excesso de características (e.g., 34.671 no SRM) prejudica o desempenho do classificador [La Croix et al. 2024]. A introdução do aprendizado profundo (deep learning, DL) marcou uma mudança de paradigma, unificando a extração de características e a classificação em um único processo otimizado e de ponta a ponta, alcançando resultados muito superiores [La Croix et al. 2024]. Diante dessa transformação, o principal objetivo do trabalho é oferecer uma pesquisa extensa e cronológica, investigando os recentes desenvolvimentos em esteganálise via DL, analisando em profundidade as técnicas e arquiteturas propostas, e identificando as tendências e os desafios emergentes, com um foco particular em imagens no domínio espacial, a fim de prover um guia valioso para a comunidade de pesquisa [La Croix et al. 2024].

A metodologia empregada pelos autores é a de uma revisão sistemática da liter-

atura, baseada no rigoroso protocolo PRISMA. Eles selecionaram 24 artigos de ponta, publicados entre 2014 e 2023, que representam a vanguarda da esteganálise com DL [La Croix et al. 2024]. O artigo estrutura-se de forma didática, iniciando com uma taxonomia detalhada das técnicas de esteganálise (dividindo-as em abordagens de assinatura e estatísticas, e subdividindo-as em específicas e universais), o que estabelece um sólido fundamento conceitual [La Croix et al. 2024]. Em seguida, a pesquisa explora a transição do paradigma de ML para o de DL, enfatizando como o DL supera as limitações do ML ao permitir uma comunicação de retroalimentação entre as fases de extração e classificação, otimizando os parâmetros de forma conjunta [La Croix et al. 2024]. O cerne da análise se debruça sobre os 24 trabalhos selecionados, dissecando as arquiteturas de Redes Neurais Convolucionais (CNNs) propostas, desde as pioneiras até as mais recentes. Essa análise detalha os componentes cruciais de cada modelo, como as camadas de pré-processamento (e.g., uso de bancos de filtros do SRM), as funções de ativação (ReLU, TanH, e a especializada TLU), as técnicas de normalização (como Batch Normalization) e as estratégias de pooling (com preferência pelo Average Pooling para preservar o fraco sinal esteganográfico) [La Croix et al. 2024].

Em vez de apresentar um novo experimento, o artigo sintetiza e analisa os resultados experimentais da literatura revisada, oferecendo uma visão panorâmica da evolução do desempenho. A pesquisa documenta a progressão das arquiteturas, começando com modelos como Qian-Net, que foi o primeiro a usar uma abordagem supervisionada com filtros passa-alta, passando pelo Xu-Net, que introduziu melhorias como a função de ativação ReLU e uma camada de valor absoluto (ABS), e chegando a modelos mais sofisticados como o Ye-Net, que inovou ao usar bancos de filtros do SRM como camada de pré-processamento e a função de ativação Truncated Linear Unit (TLU) [La Croix et al. 2024]. A análise culmina em arquiteturas estado da arte como a SRNet, uma rede residual profunda que minimiza o uso de heurísticas, e a GBRAS-Net, que se destaca pelo uso de convoluções separáveis e conexões residuais (skip connections) [La Croix et al. 2024]. Os resultados comparativos, compilados em gráficos, mostram consistentemente que o erro de detecção diminui com o aumento da quantidade de dados inseridos (payload) e que modelos mais recentes como a GBRAS-Net alcançam os menores erros, com 9.3% contra o algoritmo WOW a 0.4 bpp no dataset BOSSBase 1.01 [La Croix et al. 2024].

Um dos pontos fortes mais significativos desta revisão é a identificação e detalhamento dos principais desafios que ainda persistem no campo da esteganálise com DL. Os autores elencam sete grandes obstáculos: 1) a vulnerabilidade a ataques adversariais, que podem enganar os classificadores com modificações mínimas; 2) a qualidade e padronização dos datasets, pois a dependência excessiva do BOSSBase pode levar a resultados superotimizados e com baixa capacidade de generalização; 3) a ineficiência para lidar com imagens de tamanhos arbitrários, já que a maioria dos modelos exige redimensionamento, o que pode destruir o sinal esteganográfico; 4) a grande dificuldade na detecção de baixo payload (e.g., ¡2%), onde o sinal é extremamente sutil; 5) o problema do cover-source mismatch, que causa uma queda drástica de desempenho quando o modelo é treinado com imagens de uma fonte (e.g., uma câmera) e testado em outra; 6) a dificuldade na identificação e aprendizado de características globais, pois as operações de pooling podem diluir a informação; e 7) a necessidade de um grande número de amostras de treinamento, que são difíceis de obter [La Croix et al. 2024].

Em contraste, a proposta desenvolvida neste trabalho busca explorar o uso de bibliotecas de software livre para a construção de modelos de inteligência artificial em esteganálise.

Em contraste, a proposta desenvolvida neste trabalho busca explorar o uso de bibliotecas de software livre para a construção de modelos de inteligência artificial em esteganálise.

Em contraste, a proposta desenvolvida neste trabalho busca explorar o uso de bibliotecas de software livre para a construção de modelos de inteligência artificial em esteganálise.

Diferente de [Chubachi 2020], que enfatiza ganhos técnicos com arquiteturas avançadas, nossa abordagem objetiva democratizar a área ao garantir reprodutibilidade, baixo custo e acessibilidade, sem deixar de buscar acurácia.

3. Metodologia

4. Resultados e discussões

5. Conclusão

Referências

Chubachi, K. (2020). An ensemble model using cnns on different domains for alaska2 image steganalysis. *IEEE International Workshop on Information Forensics and Security (WIFS)*.

Foundation, F. S. (2024). What is free software? `https://www.gnu.org/philosophy/free-sw.html`. Acesso em: 27 ago. 2025.

Fridrich, J. (2010). *Steganography in Digital Media 'Principles', Algorithms, and Applications*. Springer.

Guardian, T. (2025). Apple pulls encrypted icloud storage from uk after government demands back door access. Acesso em: 26 ago. 2025.

La Croix, N. J. D., Ahmad, T., and Han, F. (2024). Comprehensive survey on image steganalysis using deep learning. 22.

Poder360 (2025). Europa barrou 41,4 mi de posts via usuários no 1º semestre de 2025. Acesso em: 26 ago. 2025.

Team, P. (2025a). About pytorch. `https://pytorch.org/projects/pytorch/`. Acesso em: 27 ago. 2025.

Team, T. (2025b). About tensorflow. `https://www.tensorflow.org/about/`. Acesso em: 27 ago. 2025.

Xu, G., Wu, H.-Z., and Shi, Y. Q. (2016). Ensemble of cnns for steganalysis: An empirical study. In *Proceedings of the 8th ACM Workshop on Information Hiding and Multimedia Security*, pages 5–10.

Yousfi, Y. and Fridrich, J. (2020). An intriguing struggle of cnns in jpeg steganalysis and the onehot solution. *IEEE Signal Processing Letters*, pages 830–834.

Yousfi, Y., Fridrich, J., Butora, J., and Tsang, F. C. (2021). Improving efficientnet for jpeg steganalysis. In *Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Secur*, pages 149–157.