Robust Steganalysis of LSB-Embedded Malicious Content Using Deep CNN

Sanjay Seshadri
Department of CSE
PES University
Bangalore, India
sanjay10.seshadri@gmail.com

Nethra Khandige
Department of CSE
PES University
Bangalore, India
nethrakhandige@gmail.com

Prof. Preet Kanwal
Associate Professor, Dept. of CSE
PES University
Bangalore, India
preetkanwal@pes.edu

***Abstract*—In the present digital age, the transmission of information via hiding it in images known as steganography has become increasingly sophisticated with new steganography algorithms entering the picture. Therefore, the need of the hour is to develop techniques to detect the presence of hidden information. As there are many steganographic algorithms in use and the list keeps growing with time, it is crucial to help detect the presence of information in images irrespective of the method used for steganography. This project presents a novel approach to blind steganalysis using neural networks, specifically focusing on deep convolutional neural networks to detect the presence of steganographically embedded malicious payloads, such as JavaScript, HTML, PowerShell, URLs, and ethereum addresses, in images without prior knowledge about the embedding algorithm. Experimental results demonstrate the effectiveness of this approach, outperforming existing implementations. This research highlights the importance of neural networks in cybersecurity and the detection of various malicious payloads in steganographic images.**

***Index Terms*—Convolutional Neural networks, Deep learning, Steganography, Steganalysis, LSB-embedding,**

I. Introduction

In the present age of digital communication, the concept of embedding information in multimedia content, mainly images, known as steganography, has seen rapid significant advancements. While steganography has its advantages, such as copyright protection and secure communication, it can also be misused for malicious activities like unauthorized data transfer and cyber espionage. The field of steganalysis, which aims to detect hidden information, has gained paramount importance in cybersecurity.

Traditional steganalysis techniques rely on manually selected features and statistical methods, which may fail to generalize over diverse datasets and steganographic methods. Deep convolutional neural networks (CNNs) offer a different approach to this problem by automatically learning hierarchical feature representations from data.

In this project, a novel deep learning framework for steganalysis using deep convolutional neural networks (CNNs) has been proposed. It leverages the power of CNNs to learn intricate features of images and detect the presence of embedded malicious payloads, such as JavaScript, HTML, PowerShell, URLs, and ethereum addresses. These payloads are embedded using the Least Significant Bit (LSB) technique. By training the network with a comprehensive dataset, a robust model is developed capable of accurately identifying steganographic content with different types of malicious payloads. This research highlights the importance of neural networks in enhancing cybersecurity through the effective detection of various steganographic threats.

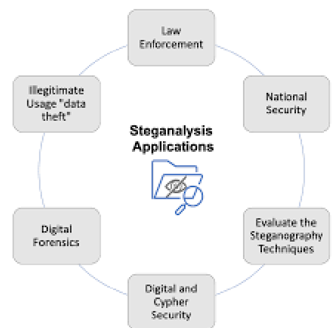

Fig. 1: Applications of Steganalysis

The information flow in the paper goes as follows, Section 2 reviews the existing approaches and implementations along with the limitations in the field of steganalysis. Section 3 presents the methodology and the architecture of the model. In Section 4, we discuss the experimental results, showcasing the effectiveness of the approach. Finally Section 5 concludes the paper and outlines the potential direction for future work.

II. LITERATURE REVIEW

The paper[1] presents a novel approach to the steganalysis of digital images utilizing convolutional neural networks (CNNs). The authors propose a unified framework that effectively integrates the essential steps of steganalysis residual computation, feature extraction, and binary classification and it allows for the direct learning of hierarchical representations from raw images.

The authors [2] propose a novel Convolutional Neural Network (CNN) architecture tailored for image steganalysis. The model takes absolute value of feature map elements of first layer to improve statistical modelling of the next layers. It also uses 1X1 convolutions as the layers get deeper. The results from the paper suggest that well-designed CNNs have the potential to provide better detection performance in the future.

According to Boroumand et al. in [6], the SRNet architecture has a deep residual design that is quite good at spotting steganography in JPEG and spatial-domain images. Modern detection accuracy is attained by this architecture with the least amount of dependence on heuristics and outside components.There are three main sections to the architecture, namely, front Segment: This segment, which consists of layers 1 through 7, is in charge of taking noise residuals out of the input images. Most notably, it stays away from using pooling layers to stop the steganographic signal from being suppressed. Middle Segment: This section, which consists of layers 8 through 12, concentrates on minimizing the feature maps' dimensionality in order to efficiently condense the collected features for additional processing.Last Segment: To differentiate between steganographic and non-steganographic images, a linear classifier called the softmax classifier comes after a fully connected layer. SRNet improves the detection capability of the model by boosting its capacity to detect subtle steganographic signals by considerably extending the front region of the detector and removing pooling layers.

The experiments, that were conducted according to the paper [8], were conducted in DCTR, GFR, and PHARM feature spaces with the aim of extracting features with and without steganographically hidden data from image pairings. Using the DCTR approach, DCT residuals were examined by convolving each block with a random 8x8 filter that was applied to the entire dataset. DCT was used to convert the resulting 8x8 residual blocks into the frequency domain. To find trends or anomalies, histograms of the DCT coefficients were made. Similarly, PHARM extracted features using a phase-aware projection model, whereas GFR used Gabor filters rather than random filters.

Both deep learning techniques and shallow machine learning classifiers in [8] were applied to the classification process. The parameter extractor-paired ensemble classifiers outperformed the latest deep learning-based systems in terms of performance. Models like XuNet, ResNet, DenseNet, and AleksNet were commonly used for deep learning, using derived picture parameters based on decompressed DCT values. With activation functions like ReLu, Tanh, TLU, sigmoid, and gaussian, these models combined convolutional, normalization, and dense layers. This allowed them to efficiently use the features that were extracted to improve the accuracy of steganalysis.

III. METHODOLOGY

The aim of the study is to develop a robust method based on deep learning to detect malicious content embedded in images by utilizing the Least Significant Bit (LSB) steganography technique. The suggested model uses a deep convolutional neural network (CNN) to identify embedded information effectively thereby amplifying the detection capabilities for cybersecurity applications.

(a) Stegano image (b) Cover Image

Fig. 2: Comparison between Stegano and Cover images.

A. Overview

The dataset [] has been carefully chosen as to suit the problem statement. The images in the dataset are then preprocessed using simple techniques such as normalizing and reshaping along with other data augmentation methods such as zooming and flipping. The preprocessed images are then fed into the neural network so as to train the model. The architecture of the model has been designed to help identify if the image being passed contains hidden information of any kind payload. The trained model is then made to test on unseen images and the model is evaluated.

B. Dataset

The dataset [7] used constitutes of 44,000 images with the resolution of each being 512x512 pixels. These images contain malicious payloads namely being JavaScript, HTML, PowerShell, URLs, and Ethereum Addresses embedded via the LSB technique. The dataset is separated into training, validation, and test sets to facilitate extensive model training and assessment. The original dataset showcased class imbalance with the cover class having one third the number of images of that of stego class. For each image in the cover class, there were three corresponding stego images with different payloads. To balance this dataset before training the model on it, each cover image was duplicated such that each image in the stego directory has a corresponding cover image. This ensures equal number of images in both classes and hence facilitating balanced training and evaluation.

C. Data Preprocessing

The images in the dataset [7] were preprocessed and hence making them ready to be used as input to the model. The pixel values were normalized to ensure faster convergence during training. Data augmentation techniques such as zooming and flipping were applied to diversify the training data and hence improve the model's generalization capabilities and reduce overfitting.

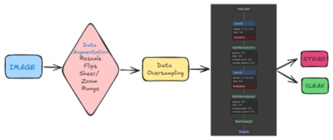

Fig. 3: Model Architecture

D. Model Architecture

The proposed CNN model is tailored to extract and learn features from images through multiple convolutional and pooling layers. The architecture is as seen below:

Convolutional Layers:

- Initial Layers: it is constituted of two convolutional layers each with 64 filters, ensued by batch normalization, pooling, and dropout. These layers capture spatial hierarchies and low-level features in the images.
- Intermediate Layers: it comprises of Two convolutional layers with 128 filters each, with comparable batch normalization, pooling, and dropout applied to augment feature extraction and minimize overfitting.
- Advanced Layers: it consists of two convolutional layers each with 256 filters, continued by batch normalization, pooling and dropout to record details and patterns of higher complexity.
- Deep Layers: it comprises of two convolutional layers with 512 filters each, designed to assimilate high-level features from the images.

Pooling and Regularization:

- MaxPooling2D: it is applied after every set of convolutional layers primarily to lessen the spatial dimensions and retain key features.
- Dropout: it is incorporated after pooling layers mainly to avert overfitting by randomly assigning a fraction of input units to zero during training.
- GlobalAveragePooling2D: assimilates the feature maps into a single vector thereby capturing global features and reducing the dimensionality from the entire image.

Fully Connected Layers:

- Dense Layers: the global pooling layer is followed by a dense layer with 512 units and ReLU activation capturing complex feature interactions. This is then followed by a dropout layer to regularize the model even further.
- Output layer: it is a final dense layer with a sigmoid activation function which is used for binary classification and determination of embedded malicious content in an image.

IV. RESULTS

The steganalysis model was trained on the dataset and had achieved upto 85% accuracy after training it on a sufficient number of epochs along with hyperparameter tuning and validation set. The number of epoch was chosen carefully so as to not result in overfitting. The model's performance was further validated on a separate test dataset which was not seen during the training phase. The model achieved a remarkable test accuracy of 95% showcasing the model's robustness in detecting malicious content in images.

V. CONCLUSION

The results of this research demonstrate that the model developed is highly effective in the task of steganalysis on LSB-embedded malicious content in images. The significant accuracy validates the robustness and generalization capability of the approach used, making it a promising solution for real-world applications in cybersecurity.

VI. FUTURE SCOPE

Although our deep CNN model has demonstrated remarkable result in the steganalysis of LSB-embedded malicious content in images, there exist multiple opportunities for further investigation and enhancement. Neural networks can be manipulated to achieve the specific task at hand. With the number of steganography techniques increasing, it is the need of the hour to be able to detect the spread of malicious content in multimedia.

REFERENCES

[1] Ye, J., Ni, J., and Yi, Y., "Deep Learning Hierarchical Representations for Image Steganalysis," IEEE Transactions on Information Forensics and Security, vol. 12, pp. 2545-2557, 2017.

[2] G. Xu, H.-Z. Wu, and Y.-Q. Shi, "Structural Design of Convolutional Neural Networks for Steganalysis," in IEEE Signal Processing Letters, vol. 23, no. 5, pp. 708-712, May 2016, doi: 10.1109/LSP.2016.2548421.

[3] N. Cassavia, L. Caviglione, M. Guarascio, G. Manco, and M. Zuppelli, "Detection of Steganographic Threats Targeting Digital Images in Heterogeneous Ecosystems Through Machine Learning," Institute for High Performance Computing and Networking (ICAR), National Research Council of Italy (CNR), Rende, Italy, pp. 1-8, June 2022.

[4] W. You, H. Zhang and X. Zhao, "A Siamese CNN for Image Steganalysis," in IEEE Transactions on Information Forensics and Security, vol. 16, pp. 291-306, 2021, doi: 10.1109/TIFS.2020.3013204.

[5] R. Zhang, F. Zhu, J. Liu and G. Liu, "Depth-Wise Separable Convolutions and Multi-Level Pooling for an Efficient Spatial CNN-Based Steganalysis," in IEEE Transactions on Information Forensics and Security, vol. 15, pp. 1138-1150, 2020, doi: 10.1109/TIFS.2019.2936913.

[6] M. Boroumand, M. Chen and J. Fridrich, "Deep Residual Network for Steganalysis of Digital Images," in IEEE Transactions on Information Forensics and Security, vol. 14, no. 5, pp. 1181-1193, May 2019, doi: 10.1109/TIFS.2018.2871749

[7] https://www.kaggle.com/datasets/marcozuppelli/stegoimagesdataset

[8] M. Płachta, M. Krzemień, K. Szczypiorski, and A. Janicki, "Detection of Image Steganography Using Deep Learning and Ensemble Classifiers," Electronics, vol. 11, no. 10, p. 1565, 2022. [Online]. Available: https://doi.org/10.3390/electronics11101565.