Image Steganalysis

Aditya Sharma A20516668
Raj J. Shah A20524266
Rutvik Savaliya A20524267

1. Introduction

1.1. Abstract

The major goal of this study is to analyze images and to identify secret messages within image-based data (2D) using the idea of *steganalysis*. As the sheer volume of data escalates annually, concerns regarding its security and dissemination intensify. Any file downloaded today can possibly contain some form of secret message that is invisible to the naked eye. Drawing parallels with cryptanalysis, steganalysis emerges as a pivotal tool for comprehending and detecting these surreptitious messages or data concealed within diverse media forms. Steganalysis defines the science behind hiding and detecting covert messages or data embedded in other forms of media. Our study aims to understand and predict the likelihood of an image containing secret data. While traditional methods rely on a foundational understanding of steganographic techniques, the infusion of machine and deep learning heralds a paradigm shift, Leveraging cutting-edge models and architectures, we aim to assess the discernment of whether an image carries a covert message and, if feasible, identify the steganographic algorithm responsible for embedding this secretive information.

1.2. Steganography/Steganalysis

This secret art of concealing information within seemingly innocuous media, known as *steganography*, has evolved alongside its counterpart, *steganalysis* – the art of deciphering these embedded secrets.

The term ”steganography” originates from Greek, a fusion of ”steganos” meaning ”covered” and ”graphein” meaning ”writing”. It encapsulates the essence of embedding messages within obvious carriers, such as images, audio, or video, without altering their apparent form.

Steganography and steganalysis represent an interplay between concealment and detection, a continuous dance between those who seek to hide and those who seek to find. In the realm of image steganalysis, we delve into the intricate details of digital images, identifying binary patterns and pixel relationships to uncover hidden messages.

Steganalysis techniques can be designed for specific embedding algorithms, deciphering the modifications introduced by these methods. Alternatively, general techniques can be employed to detect any form of hidden data, regardless of the underlying algorithm.

Steganalysis is an intricate endeavor. Its complexity stems from the inherent limitations of the cover media used to conceal these clandestine communications. Over the years, image-based steganography has emerged as a prevalent approach, yet its very prevalence has fueled the development of sophisticated steganalysis techniques. *Any steganographic technique can be exposed once we understand what steganalysis technique was used. [1]*

There are several steganographic techniques developed over decades, that must satisfy at least these three conditions:

1. The maximum capacity that can be embedded should be within the limits of the ”cover” image.
2. The quality(visually) of the embedded (or stego) image must be similar to the ”cover” image.
3. It should be robust to noise.

All of the embedding techniques can be categorized into spatial or transform domains. ’Spatial’ based

techniques are simpler because they focus on the actual location of the pixel of information within the image but are less robust, especially to noise. 'Transform-based' techniques are used which can transform the signal into another form while retaining the information. Wavelet transform— is one such technique that can transform 'spatial' into 'frequency' based.

In our study, we focus on using the ALAKSA2 Dataset [2], developed by Troyes Institute of Technology, consisting of providing 'stego' images using 3 steganographic algorithms, along with the cover images.

1. J-UNIWARD (JPEG - universal wavelet relative distortion) [3]: A 'Transform' based algorithm that focuses on the 'wavelet' transform of an image that, unlike traditional steganographic algorithms that modify LSBs uniformly, UNIWARD employs a more selective strategy (relative to a pixel position, correlation), focusing on pixels with higher visual importance and lower embedding capacity.

2. J-MiPOD (JPEG - Minimizing Performance of Optimal Detector) [4]: A 'Spatial' based algorithm that focuses on the 'Middle' Significant Bits of the image. It utilizes a novel embedding strategy that leverages the statistical properties of the image to minimize the distortion caused by embedding. While MiPOD is considered a spatial-domain algorithm, it does incorporate some elements of transform-domain steganography. Specifically, it utilizes an estimated MVG (Multivariate Gaussian) model to characterize the statistical distribution of MSBs in the image.

3. UERD (Uniform Embedding Revisited Distortion) [5]: A 'Spatial-based algorithm similar to UNIWARD focuses on the Least Significant Bits within an image, by taking into account the spatial correlation of neighboring pixels when determining the 'embedding efficiency' of a pixel. Pixels with high embedding efficiency are less susceptible to distortion and are therefore selected for embedding secret bits. To determine the embedding efficiency of a pixel, UERD considers the local spatial correlation of neighboring pixels. Areas with high spatial correlation, such as smooth regions or textureless backgrounds, are more tolerant of LSB modifications

1.3. Related Work

One major difference between cryptography and steganography is the idea that cryptography 'hides' the message, whereas steganography hides the 'presence' of the message. Studying and Analysing the current techniques for image-based steganalysis leads to better results in steganography, If we can decipher these techniques, it becomes possible to unveil the concealed information within an image. The Passive Steganalysis technique relies on just detecting the existence of hidden messages within images, compared to Active Steganalysis which retrieves the hidden messages. Many methods have been proposed for steganalysis applications based on machine learning and deep learning algorithms.

A standard machine-learning technique for steganalysis involves two parts. First, is the *feature extraction* followed by the *classifier*. The features extracted from the images are used as input for the classifier. Since it can be treated as a classic 'classification' or 'regression' problem, we can utilize techniques like linear regression, PCA, K-Nearest Neighbour, or K-Means Clustering.

In the era of burgeoning computing power and the proliferation of Graphics Processing Units (GPUs), the landscape of steganalysis went through a shift, finding relevance in deep learning techniques. A moment materialized in 2015 [6] with the adoption of Convolutional Neural Networks (CNNs) as the primary architectural choice. What renders this paradigm particular is its seamless integration of both feature extraction and classification within a singular black-box entity. This stands in contrast to conventional machine learning approaches, wherein the classifier's efficacy is dependent upon the proficiency of feature extraction.

The distinctive advantage of employing CNNs lies in their intrinsic ability to autonomously learn and discern relevant features, obviating the need for explicit feature engineering. This marks a departure from traditional methods, where the performance of the classifier was intricately entwined with the quality of manually crafted features. To avoid the curse of dimensionality, several techniques have been proposed for fea-

ture selection such as PCA, employed by Gorkar [7] as well as Desai [8] on the eigenvalues.

For spatial embedding-based techniques, the Spatial Rich Models (SRM) developed by Fridrich [9] stand as widely adopted, image steganalyzers. SRM derives residual features by applying both nonlinear and linear high-pass filters. Qian [6] was the first to propose using supervised learning with CNNs for steganalysis. Their network consists of three steps, a high-pass filter used as a preprocessing layer, a convolutional layer for feature extraction and then a fully connected layer for classification. Liu [10] introduced features extracted from the Subtraction Pixel Adjacency Model (SPAM) in conjunction with the binary bat method for feature subset selection. Xu [11] proposed utilizing a Convolutional Neural Network (CNN) to circumvent network convergence issues. Their approach involved using a high-pass filter as a layer to capture noise residuals from original images, followed by five convolution and pooling layers. Zeng [12] employed a hybrid deep learning model, integrating techniques such as quantization and truncation.

2. Methodology

Our proposed methodology revolves around identifying techniques and employing deep learning models like Convolutional Neural Networks (CNNs), Residual Networks (ResNets), and other state-of-the-art models. We also utilize data augmentation and transformation techniques to develop a method capable of classifying which steganography method has been employed (among the three available in the dataset) and detecting the presence of concealed images within an image. Additionally, we explore the feasibility of Generative Adversarial Networks (GANs) and other transformer-based techniques, while striving to establish more generally applicable approaches that can be deployed on any range of images.

In our analysis, we have utilized standard steps in traditional machine/deep learning process along with Exploratory Data Analysis :

1. Data Preprocessing/Acquistion: We will begin by collecting and preparing relevant data from Kaggle for our deep learning model's training. This meticulous process ensures the model receives high-quality data for optimal learning.

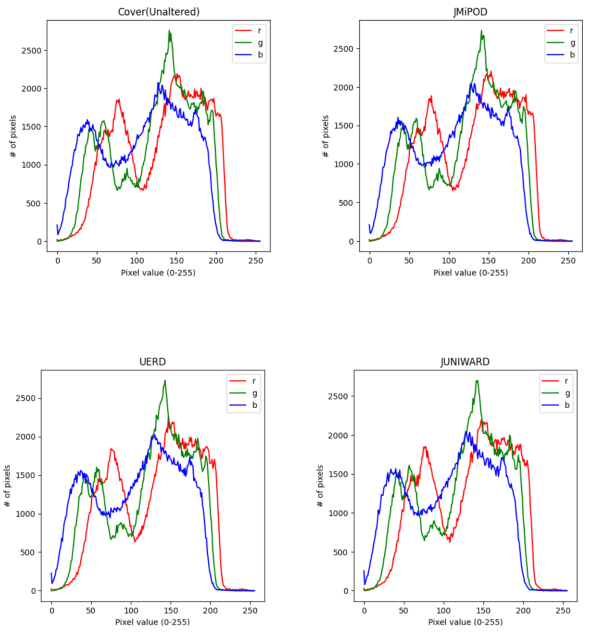
2. Exploratory Data Analysis: We will delve deep into the collected dataset through an in-depth exploratory data analysis (EDA). This comprehensive analysis will enable us to extract valuable insights and patterns, differentiate between images, and understand the distribution across the three defined steganographic algorithms.

3. Model Training: Armed with the insights gleaned from the EDA, we will proceed to train our chosen state-of-the-art models. Utilizing Tensor Processing Units (TPUs) for efficient computation, we will define the parameters of these models to achieve optimal performance.u

4. Hyperparameter Tuning: We will rigorously evaluate our models and tune their hyperparameters to achieve the best training parameters. This deliberate optimization process, facilitated by the effective "Early Stopping" and "Reduce Learning Rate on Plateau" callbacks, ensures our model achieves better validation accuracy while avoiding overfitting, leading to enhanced performance.

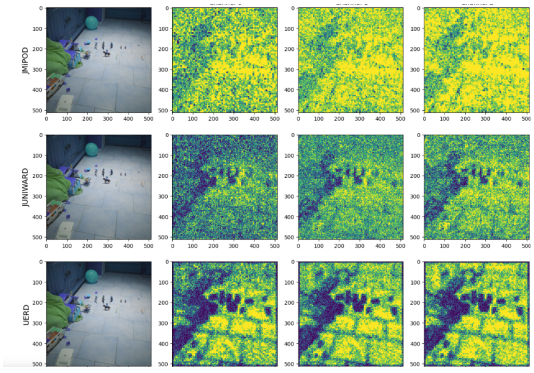
5. Visualizations: To effectively communicate our findings and showcase the results of our analysis and modeling, we will leverage impactful data visualizations.

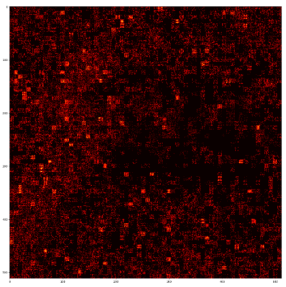
Pixel Distribution (RGB) across the cover as well as stego-algorithms

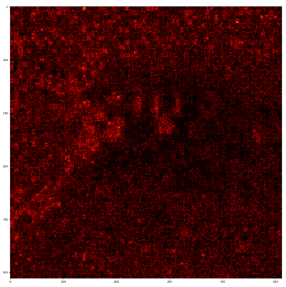


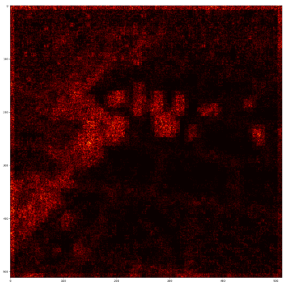
Harnessing the capabilities of deep learning, we employed cutting-edge image classification and detection models to address the critical challenge of identifying hidden data within images. By harnessing the capabilities of robust AI hardware and GPUs, our approach achieves exceptional performance in uncovering hidden information, unlocking new possibilities and paving the way for advancements across various domains, especially with larger datasets.

Channel-Wise Distribution of a single image across the 3 stego-algorithms









The Different Model Architectures implemented in our Analysis:

- **SRNet** [13]: The Structure-Responsive Network represents a formidable neural network designed for handling of tasks such as image classification and text editing within images. Employing a twin discriminator Generative Adversarial Network (GAN) architecture, the model utilizes convolutional layers to extract intricate visual features, including edges, patterns, and textures. An attention block is incorporated to selectively concentrate on sections within the image, optimizing task-specific focus. The adaptive nature of the GAN architecture equips the model to excel across diverse tasks, particularly demonstrating exceptional performance on extensive datasets.

- **EfficientNet** [14]: EfficientNet stands as a robust convolutional neural network known for its ”compound” scaling strategy, which uniformly scales images across all dimensions through a singular coefficient. Distinguished by its architectural components such as ”depthwise” separable convolutions, this network effectively mitigates computational costs while dynamically adapting channel weights based on their significance. Notably, the incorporation of advanced features like ”AutoAugment” enhances training efficiency by autonomously identifying ideal augmentation techniques. EfficientNet is further exemplified by an array of pre-trained models, making it adaptable to diverse hardware constraints and ensuring user-friendly implementation.

- **InceptionResNet** [15]: The InceptionResNet represents a powerful convolutional neural network (CNN) architecture, seamlessly integrating the strengths of two state-of-the-art structures,

namely Inception and ResNet. At its foundation, the network relies on Inception blocks as fundamental components, deploying parallel filters to capture spatial and frequency patterns inherent in the image data. Noteworthy is the incorporation of residual connections within these Inception blocks, strategically implemented to prevent challenges such as the vanishing gradient problem and facilitate the learning of more complex features. Demonstrating exceptional proficiency in tasks related to image recognition and classification, this architectural innovation allows for the construction of deeper networks, effectively addressing concerns such as the vanishing gradient phenomenon..

- **MobileNet** [16]: MobileNet represents a convolutional neural network (CNN) architecture characterized by its lightweight design, specifically tailored for deployment on compact embedded and mobile devices. Employing the approach known as "depthwise" separable convolution, MobileNet mitigates computational demands. Additionally, it incorporates diminutive "bottleneck" layers to effectively reduce both parameters and computations. This crafted architecture is particularly well-suited for resource-constrained environments, catering to the unique demands of small-scale devices. Furthermore, its adaptability extends to customization for diverse applications and hardware configurations.

3. Results

Following the completion of our experiment involving various models for detecting hidden data within images, it is evident that the EfficientNet model demonstrated superior accuracy compared to other models. This trend persisted when evaluating the testing dataset.

Subsequent to the EfficientNet model, the InceptionResNet exhibited commendable performance, attributed to its larger architecture featuring residual connections.

Contrastingly, state-of-the-art models such as SRNet and MobileNet displayed suboptimal performance and susceptibility to overfitting, resulting in lackluster results on the testing set. To mitigate the need for re-training and optimize computational resources, we strategically saved checkpoints based on the best performance for each model. Furthermore, we implemented a reduced batch size of 16 to enhance overall performance. However, it is imperative to note that solely relying on accuracy might not provide a comprehensive understanding of model performance. Considering the intricacies of the task at hand, we recommend supplementing accuracy with additional evaluation metrics, such as Area Under the Curve (AUC) or Weighted AUC, to gain a more nuanced perspective.

	SRNet	EfficientNet(B7)	InceptionResNet(v2)	MobileNet(v2)
Training Acc.	49.62	66.81	50.14	49.69
Validation Acc.	49.84	52.81	50	49.98
Test Acc.	48.05	51.85	50.8	51.28

Table 1. Accuracy over different model architectures

4. Conclusion

Steganalysis, the art of detecting hidden data embedded within digital media, is a rapidly evolving field with far-reaching implications for data security and privacy. Our exploration of steganalysis techniques, focusing on three specific algorithms, underscores the intricate nature of hidden data detection and identification. We conclude that effectively uncovering hidden data requires a deep understanding of underlying algorithms and domain knowledge, often necessitating the development of tailored architectures specific to the employed method.

While our research has yielded valuable insights, we acknowledge the potential for further refinement, particularly in enhancing model performance. This refinement can be achieved through hyperparameter tuning and the incorporation of larger datasets. By employing a diverse range of models and architectures, we have gained a comprehensive understanding of their respective strengths and limitations within the context of image classification, detection, and large-scale image processing. This knowledge enables us to make informed decisions regarding the suitability of specific models and their corresponding hyperparameters.

4.1. Future Aspirations

Based on the results of our analysis and the results of the dataset, there are possible avenues for future re-

search:

- **Feature Extraction**: The accuracy of steganalysis hinges on the ability to extract meaningful and distinctive features from media. Exploring novel feature extraction techniques and leveraging pretrained features can significantly enhance model performance.

- **Computational complexity**: Steganalysis, particularly when dealing with image-based data, can be computationally demanding, especially for large datasets. This limitation hinders its real-time applicability. Addressing this challenge requires developing more efficient algorithms and leveraging hardware acceleration techniques.

- **Performance using other models**: Our analysis primarily focused on a limited subset of state-of-the-art models. Incorporating a wider range of image detection and classification models, along with developing architectures for our specific scenario, holds immense potential for performance improvements. Additionally, techniques like ensembling and hybrid frameworks can further enhance model capabilities.

- **Evaluation metrics**: The prevalent focus on 'detection accuracy' in steganalysis evaluation may not fully capture the essence of steganalysis performance. Employing more comprehensive metrics, such as false positive or negative rates, ROC curves, can provide a more holistic evaluation.

- **Generalization over Steganographic Algorithms**: The limitation of our dataset to only three algorithms restricts the 'standardization'. Expanding the dataset to incorporate a broader range of steganographic algorithms will enable the development of a universal 'hidden' data detection system for images, capable of crossing dimensional barriers (e.g., 3D images)

5. Source Code & Data

- Data: Kaggle

- Source Code: Github

6. Acknowledgment

We would like to express our sincere gratitude to *Prof. Yan Yan* for providing us with guidance and support throughout this project. His teachings, insights, and feedback have been invaluable in shaping the final project and report.

References

[1] Wafa M. Eid, Sarah S. Alotaibi, Hasna M. Alqahtani, and Sahar Q. Saleh. Digital image steganalysis: Current methodologies and future challenges. *IEEE Access*, 10:92321–92336, 2022.

[2] Rémi Cogranne, Quentin Giboulot, and Patrick Bas. The alaska steganalysis challenge: A first step towards steganalysis. In *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security*, IHMMSec'19, page 125–137, New York, NY, USA, 2019. Association for Computing Machinery.

[3] Vojtech Holub, Jessica Fridrich, and Tomáš Denemark. Universal distortion function for steganography in an arbitrary domain. *EURASIP Journal on Information Security*, 1, 12 2014.

[4] Vahid Sedighi, Rémi Cogranne, and Jessica Fridrich. Content-adaptive steganography by minimizing statistical detectability. *IEEE Transactions on Information Forensics and Security*, 11(2):221–234, 2016.

[5] Linjie Guo, Jiangqun Ni, Wenkang Su, Chengpei Tang, and Yun-Qing Shi. Using statistical image model for jpeg steganography: Uniform embedding revisited. *IEEE Transactions on Information Forensics and Security*, 10(12):2669–2680, 2015.

[6] Yinlong Qian, Jing Dong, Wei Wang, and Tieniu Tan. Deep learning for steganalysis via convolutional neural networks. *Proceedings of SPIE - The International Society for Optical Engineering*, 9409, 03 2015.

[7] Yogesh Kulkarni and Anurag Gorkar. Intensive image malware analysis and least significant bit matching steganalysis. 12 2020.

[8] Madhavi Desai and S.V. Patel. Pfa-based feature selection for image steganalysis. *International Journal of Bioinformatics Research and Applications*, 14:119, 01 2018.

[9] Jessica Fridrich and Jan Kodovsky. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7:868–882, 06 2012.

[10] Feng Liu, Xuehu Yan, and Yuliang Lu. Feature selection for image steganalysis using binary bat algorithm. *IEEE Access*, PP:1–1, 12 2019.

[11] Guanshuo Xu, Hanzhou Wu, and Yun Shi. Ensemble of cnns for steganalysis: An empirical study. pages 103–107, 06 2016.

[12] Jishen Zeng, Shunquan Tan, Bin Li, and Jiwu Huang. Large-scale JPEG steganalysis using hybrid deep-learning framework. *CoRR*, abs/1611.03233, 2016.

[13] Liang Wu, Chengquan Zhang, Jiaming Liu, Junyu Han, Jingtuo Liu, Errui Ding, and Xiang Bai. Editing text in the wild. *CoRR*, abs/1908.03047, 2019.

[14] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946, 2019.

[15] Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261, 2016.

[16] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017.