Research Paper #4: ‘Enhancing Steganography Detection with AI: Fine-Tuning a Deep Residual Network for Spread Spectrum Image Steganography’

Learning Outcomes:

- Create a summary of notes for reference and understanding
- Understand Steganography Detection leveraging AI
- Review trending AI models for Steganography detection
- Review knowledge of Steganography

Overview

- This paper investigates the usage of Artificial Intelligence (AI) and Convolutional Neural Networks (CNNs) in image steganography detection

> *“The advent of artificial intelligence (AI) has ushered in a new era for image steganography detection. AI-based techniques can learn to recognize the subtle changes induced by steganography, outperforming traditional methods in detecting concealed information. Deep learning, a subset of AI, has achieved impressive results, with Convolutional Neural Networks (CNNs) successfully identifying minute alterations in image structures. The*

> *adaptability of AI makes it a powerful tool against the ever-evolving tactics of data-hiding."*

- Artificial Intelligence is seen as a promising tool in Steganography Detection
- CNNs possess the ability to recognise small changes in an image and could thus counter most Steganography techniques
- AI is adaptable, meaning it will keep pace with evolving steganography tools and techniques

> *"However, AI-based image steganography detection is not without its hurdles. The reliance on vast, labeled datasets for training is a significant barrier, considering the inherent difficulty in procuring such datasets in this secretive field. The computational resources needed for such techniques are also a considerable concern. Furthermore, the 'black box' nature of these AI models makes understanding the decision-making process a challenge"*

- AI models require many labelled training data samples. Steganography makes this a challenge because it is difficult to identify a known steganographic image in the wild and subsequently create a dataset that can be used to train a model
- AI models can have high computational costs; resource-intensive
- Deep learning AI models are black box, meaning their decision-making abilities can be difficult to understand.

Steganography Embedding

A Stego image S can be formed from an image I of size $m * n$ where each pixel $p_{i,j}$ is represented by a value from a finite set Γ, corresponding to the range [0,255] for an 8-bit grayscale image

To embed the secret message M of length k within I, the function $f : \Gamma * M \rightarrow \Gamma$ is applied, taking a pixel value and a portion of the secret message and returning a new pixel for the stego image. The overall steganography process can be expressed as

$$S = f(I, M)$$

> *"It is important to note that an ideal steganography function f will ensure that S is statistically indistinguishable from I. The difficulty of detecting the existence of the secret message M within S depends on how closely the distribution of S resembles the distribution of I"*

AI-Based Steganography Detection

Convolutional Neural Networks

- CNNs are increasingly being used for image steganography detection

> *"CNNs are a type of deep learning model that are especially adept at processing grid-like data, such as images. The core principle behind CNNs is the concept of 'convolution', a mathematical operation that fuses two functions to produce a third. This operation, when applied to image processing, allows CNNs to effectively identify complex patterns in images, such as the subtle alterations induced by steganography"*

> *"The general architecture of a CNN-based steganalysis model comprises an input layer, several convolutional layers, pooling layers, fully connected layers, and an output layer. The input layer receives the image, while the convolutional layers are responsible for feature extraction. These layers apply a series of filters to the input image, detecting various features such as edges, textures, and shapes. The pooling*

> *layers perform down-sampling operations to reduce computational complexity and control overfitting."*

> *"The extracted features are then flattened and passed onto the fully connected layers, which perform high-level reasoning based on these features. Finally, the output layer makes the prediction—whether the image contains hidden information or not"*

Deep Residual Network for Steganalysis (SRNet)

- SRNet is a deep learning model specifically designed for steganalysis

> *"The SRNet comprises a total of twelve layers. The initial two layers kick-start the process of feature extraction, using 3 × 3 filters, which serve to increase the number of kernels to 64 before reducing to 16 feature maps to economize on memory consumption. Importantly, these layers do not include any pooling or residual shortcuts"*

> *"The strengths of the SRNet architecture lie in its judicious blend of convolution, pooling, and residual learning. This structure allows the network to learn both local and global features of the image, efficiently distinguishing between stego and non-stego images"*

Methodology Review

- This paper builds on top of the existing SRNet model and aims to refine its recognition for Spread Spectrum Image Steganography (SSIS)
- SSIS hides data in a spread-spectrum manner, making it resemble natural noise
- Standard CNNs (like SRNet) struggled with detecting SSIS due to its diffused and low-amplitude modifications

Modifications to SRNet:

- Adaptation of Training Strategy
  - Low initial learning rate (1×10^{-4}) to preserve pre-trained features
  - Progressive unfreezing: Start training with early layers frozen, then gradually unfreeze them
  - Batch size set to 32 to balance computational efficiency and learning stability
- Refinements in Model Architecture
- Optimisation of Training Process
  - Balanced mini-batches: Equal numbers of stego and cover images
  - Data augmentation: Random cropping and rotation
  - Early stopping based on validation loss to avoid overfitting
- Overall accuracy improved for all payload levels (bits per pixel)
  - 0.125bpp: Accuracy increased from 60.3% to 72.15%
  - 0.25bpp: Accuracy increased from 65.50% to 79.30%
  - 0.5bpp: Accuracy increased from 78.55% to 88.40%
- These performance improvements came with a trade-off, however, as it came reduced performance for detecting other steganographic techniques

Building AI Model for Steganography Detection

(Content generated by ChatGPT)

Public Datasets:

- **BOSSBase (v1.01)**
  - 10,000 grayscale cover images (512×512), commonly used in steganalysis.
- **BOWS2**
  - Similar to BOSSBase, good for generalization tests.
- **ALASKA & ALASKA2**
  - Designed for deep learning steganalysis, with multiple stego methods and JPEG images.
- **StegoAppDB**

    - Focuses on mobile steganography tools — gives real-world examples of how stegomalware could behave.
- Optional: Create your own stego images using tools like **OpenStego**, **Steghide**, **F5**, or **HUGO** to control the payload.

Model Architecture

Suggested Baseline Models:

- Xu-Net: Easy to implement, great baseline.
- Ye-Net or Yedroudj-Net: Stronger and more modern CNNs.
- SRNet: The state-of-the-art, deeper but more resource-intensive.

If you want to go big:

- Combine CNN with attention layers or transformer blocks.
- Use contrastive learning (e.g., SimCLR) for better feature learning.

References

Kuznetsov, O., Frontoni, E., Chernov, K., Kuznetsova, K., Shevchuk, R., & Karpinski, M. (2024). Enhancing Steganography Detection with AI: Fine-Tuning a Deep Residual Network for Spread Spectrum Image Steganography. *Sensors*, *24*(23), 7815–7815. https://doi.org/10.3390/s24237815

ChatGPT Q&A

https://chatgpt.com/share/67f37f4a-bee8-8001-a42f-d269ad1465d5