Steganoanalysis

Illia Kostenko

Masaryk university Knowledge discovery group

Main goals

The main goal is analyzing steganography algorithms, create changed with steganography images and use them with machine learning models for the purpose of steganalysis. Is required to create and train model, and then tune it for each of the classes.

Also, is recommended, to preprocess the images, i.e. change the quality of images, resize them. For the encoding messages in the images are used algorithms UERD, UNIWARD, UPiMOD.

It is needed to analyse changes of created, encoded images and analyse obtained results after fitting ML model and predicting basic results.

The final objective of this project is a classification task where we would be building a reliable system capable of detecting secret data within innocuous-seeming digital images.

What is steganography?

Steganography is a technique of hiding secret data in a common, non-secret, file or message to avoid detection; the secret data is then extracted to its destination. Steganography includes the concealment of information within computer files. In digital steganography, electronic communications may include steganographic coding inside of a transport layer, such as a document file, image file, program, or protocol. Main areas of steganography usage are:

- Media Database systems
- Access control system
- Protection of data alteration
- Confidential communication

The key difference between cryptography and cryptography is the use of a " key " in cryptography to convert plain text into cipher text, which would ensure that the user cannot understand the hidden message. While in steganography the information is hidden so that the structure of the covering particle does not change, and the very fact that the message is hidden is unknown. Steganalysis is the process of detecting hidden data that are using steganography. Steganography and steganalysis are analogous to encryption and decryption.

Used steganography algorithms

These algorithms primarily hide messages into the DCT coefficients of the images. A discrete cosine transform (DCT) expresses a finite sequence of data points in terms of a sum of cosine functions oscillating at different frequencies.

$$DCT(P) = \sum_{i=1}^{\infty} \frac{w_k w_l}{4} cos\frac{\pi k(2i+1)}{16} cos\frac{\pi k(2j+1)}{16} P$$

where k,l are 0,...,7 index the DCT mode and $w_0 = 1/\sqrt{2}, w_k = 1 for k > 0$.

- **JUNIWARD** The distortion function of J-UNIWARD needs to filter images three times to get HL,LH and HH subbands, and then combine three subband coefficients to calculate the embedding distortion for each element.
- **UERD** Followed by the concept in spirit of “spread spectrum communication”, UED (Uniform Embedding Distortion) and UERD (Uniform Embedding Revisited Distortion) with low complexity uniformly spread the embedding modifications to DCT coefficients of all possible magnitudes.
- **MiPOD** As in the UNIWARD algorithm, distortion function ρ associates to each pixel the cost of modifying it.

Used dataset

We will be using a dataset from a kaggle competition known as ALASKA2 Image Steganalysis.

Number of images in each folder, **Cover, JUNIWARD, UERD, JMiPOD**: 7500

Size of images ≈92 KB

Dimensions: 512x512 pixels.

Also, were created 9 directories for the preprocessing images with changing quality of the images. Directories “JUNIWARD75”, “JMIPOD75”, “UERD75” contain images, that were preprocessed with PIL library, and that quality of which was downgraded due to the coefficient 75 of PIL. Folders “JUNIWARD90”, “JMIPOD90”, “UERD90” are using coefficient 90 in PIL and folders “JUNIWARD95”, “JMIPOD95”, “UERD95” are using coefficient 95.

Used neural networks for the analysis

For the practical implementation task of steganalysis was decided to use models of CNN architecture.

- **Conv2D** Created by own model.
  Obtained accuracy for this model is 60%.
- **EfficientNet B0** beginning variant of the model.
  Obtained accuracy for this model is 61%.
- **EfficientNet B3** Further realization of a EfficientNet model.
  Obtained accuracy for this model is 65%.
- **EfficientNet B7** The latest realization of a EfficientNet model.
  Obtained accuracy for this model is 75%.

Testing steganalysis for different classes of images

Main idea was to test hypothesis of quality steganalysis for different class pictures. For the purposes of analysis was decided to use network EfficientNet B0.

Table 1:Obtained results for B0 for different classes

Table 2:

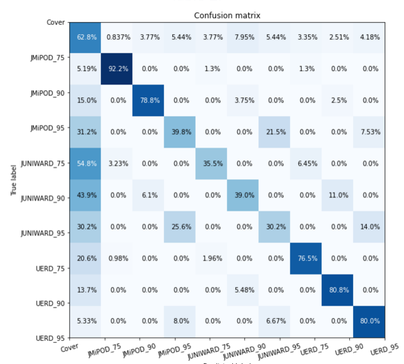


Table 3:

Classes	Accuracy %
Cover	62.7
JMiPOD75	92
JMiPOD90	78
JMiPOD95	39
JUNIWARD75	35
JUNIWARD90	39
JUNIWARD95	30
UERD75	76
UERD90	80
UERD95	80

Final accuracy is 61 perc. Weighter AUC is 0.883.

Analysis of obtained results

Analysing results of the first model Conv2D

Metrics: training loss and validation loss. Values of training loss comes down to 0.2336, accuracy is 60% and the validation loss is at 0.433 after 30 epochs.

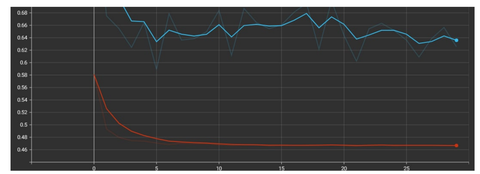


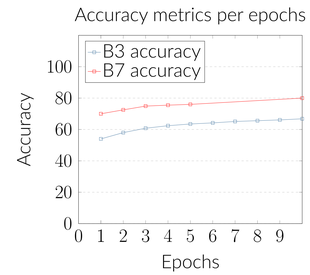
Table 4: Obtained metrics for Conv2D network

Model created from B3

EfficientNet B3 is the fourth generation of EfficientNet models, and it should be more effective. But, obtained results show us very high (comparing to another models) loss values and quite low accuracy – 68%.

Model created from B7

Metrics: accuracy, valaccuracy. Obtained valloss is 0.52 and loss is 0.51. These values are bigger than for the model in previous examples, but for used five epochs was obtained quite high accuracy – 75% per 5 epochs .



References

[1] Wenbo Zhou Weiming Zhang Nenghai Yu Kejiang Chen, Hang Zhou.
Defining cost functions for adaptive jpeg steganography at the microscale, 2019.

[2] Michael T.Raggo.
Steganography, steganalysis and cryptoanalysis.
DefCon12, 2004.

[3] Xianglei Hu Jiwu Huang Wenkang Su, Jiangqun Nia.
New design paradigm of distortion cost function for efficient jpeg steganography.
School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China, 2021.