MULTI-DATASETS EVALUATION OF GB-RAS NETWORK BASED STEGDETECTORS ROBUSTNESS TO DOMAIN ADAPTATION PROBLEM

Dmytro Progonov

***Abstract**: Protection of state and commercial critical infrastructures is topical task today. Among known threats, special interest is taken to early detection of hidden (steganographic) channels usage by attackers for unauthorized transmission of sensitive data. These channels are created by modification of the innocuous files, such as digital images, for message hiding and further transmission of modified files. Detection of formed stego files is non-trivial task due to the wide usage by attackers of adaptive embedding methods that preserve low level of cover image distortion during embedding. Modern solutions for stego images detection are based on utilization the novel convolution neural networks for revealing weak alterations of cover’s features during message hiding. Training of such networks is performed on fixed set of standard image databases, such as BOSS, ALASKA etc. Despite high detection accuracy of pre-trained stegdetectors, theirs performance “in the wild” on natural images remains an open question. The work is devoted to performance analysis of advanced GB-Ras convolutional network based stegdetector for adaptive embedding methods detection on natural images presented in various datasets. Obtained results proved the effectiveness of neural network applying for mitigation with domain adaptation problem for modern stegdetectors. The GB-Ras model allows improving detection accuracy of stego images for standard ALASKA (up to 6% for for S-UNIWARD embedding methods) and VISION (up to 13% for S-UNIWARD and to 9% for MG embedding methods) datasets in comparison with cover rich models based stegdetectors. Nevertheless, the performance of considered GB-Ras based stegdetector for high quality images from MIRFlickr dataset is conceded to maxSRMd2 rich model based stegdetector (up to 11% for S-UNIWARD and 12% for MG* *embedding methods). Therefore, considered GB-Ras network may be used as a promising candidate for further design of novel stegdetectors that are robust to domain adaptation problem.*

Keywords: *digital images steganalysis, adaptive embedding methods, convolutional neural networks.*

ACM Classification Keywords: *Security and privacy – Systems security – Information flow controls*

DOI: *https://doi.org/10.54521/ijita28-04-p04*

Introduction

Reliable protection of sensitive data processed in critical information infrastructure of state agencies as well as enterprises is one of the topical tasks today. Special interest is taken on counteraction to hidden (steganographic) communication channels usage by attackers for unauthorized sensitive data transmission as well as information sharing during attacks [Kopeytsev, 2020]. Feature of such channels is modification of innocuous files, such as digital images (DI), for message hiding and further transmission of created stego files to recipients.

Proposed solutions for revealing of steganographic channels are based on usage of stegdetectors (SD) – specialized data processing modules for revealing weak abnormal changes of cover image (CI) caused by stego data hiding [Fridrich, 2009; Konahovych et al, 2018]. Modern SD are based on usage the novel convolutional neural networks (CNN) for accurately detection of mentioned changes of CI. This makes possible to provide high detection accuracy for wide range of steganographic methods and even novel adaptive embedding methods that minimizes CI distortions during message hiding.

Despite high detection accuracy, CNN-based stegdetectors may be vulnerable to domain adaptation problem [Ker et al, 2013] – sensitivity to statistical features changes for images taken from different sources (datasets). Majority of researches in the digital image steganalysis domain does not consider this case and focus only on SD performance improving on fixed set of standard datasets.

Therefore, the question of SD performance by the working “in the wild” remains topical today.

Current work is aimed at filling this gap by evaluation of advanced CNN-based stegdetectors on several public datasets. Obtained results allow to make recommendation for further improvement of SD performance by working with natural images.

Related works

The state-of-the-art approach to design stegdetectors for digital images is based on digital image pre-processing (calibration) for cover image content suppression and further applying of rich models (RM) for features extraction from obtained residuals [Fridrich et al, 2012; Boroumand et al, 2018]. The calibration is performed by applying of extensive set of high-pass filters [Fridrich et al, 2012]. The RM is based on utilization of Markov chains for estimation correlation between adjacent elements of obtained residuals. This approach gives opportunity to achieve high detection accuracy by the cost of computation complexity and limited flexibility for SD fast adaptation to new embedding methods.

For overcoming mentioned limitations, in the work [Tan et al, 2014] it was proposed to utilize flexibility of novel CNN. Training of these networks is based on effective error backpropagation algorithm that allows learning appropriate convolutional kernels for effectively suppression of CI content instead of using of hand-crafted filters. The comparison of both mentioned pipelines is presented at Fig. 1 [Reinel et al, 2019].

The classical approach of DI steganalysis can be divided into two stages (Fig. 1, top side). The first one consists of image’s content suppression and further calculation of statistical features from obtained residuals. Modern approach to image content suppression is based on usage of filters from Spatial Rich Models (SRM) that provides near to the state-of-the-art detection accuracy [Fridrich et al, 2012]. Nevertheless, the question of an appropriate filter selection for effective content suppression for wide range of natural images remains open. The second stage is aimed at decision making for prepared features with usage of pre-trained classifier, such as Support Vector Machine, Random Forest [Kodovsky et al, 2012].

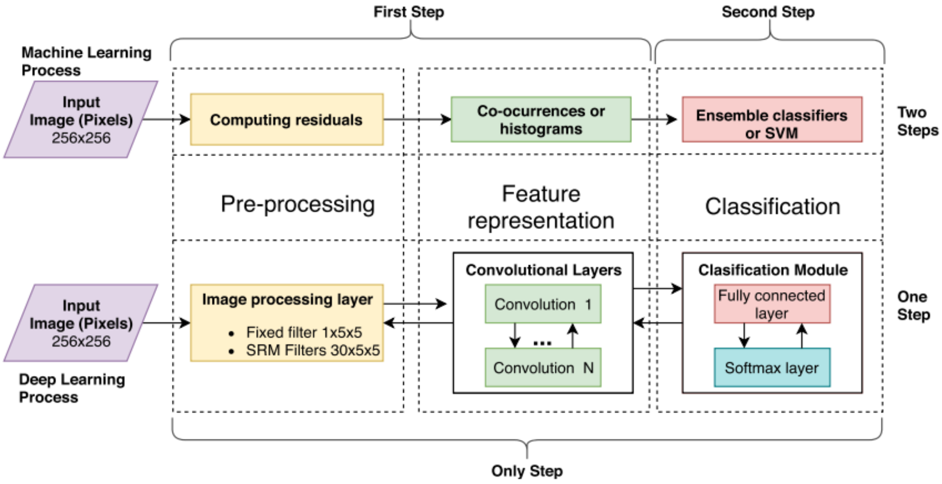


Figure 1. Comparison of the classical (top side) and deep learning (bottom side) based digital image processing pipelines in stegdetectors [Reinel et al, 2019].

The deep learning approach (Fig. 1, bottom side) allows joining both mentioned two stages in a single CNN architecture, for example ZhuNet [Zhang et al, 2018], IAS-CNN [Jin et al, 2020], DRHNet [Xu et al, 2020], SCAE [Tan et al, 2014], UT-SCA-GAN [Yang et al, 2018], etc. It gives the opportunity to accelerate the optimal calibration parameters search that minimizes stego image classification error. As an example, modern CNN-based stegdetectors may utilize SRM filters as initial convolutional kernels and further tune them during training procedure [Tan et al, 2014; Ye et al, 2017]. Also, CNNs allow including prior information about embedding methods features into training process, namely applying shortcuts, spatial pyramid pooling (SPP) [He et al, 2015], heterogeneous convolutional kernels [Singh et al, 2019], different kinds of regularization [Bisong, 2019], implementation of subnetworks ensemble [Xu et al, 2016a], for additionally improving detection accuracy.

Despite high effectiveness of proposed CNN-based stegdetectors, their performance is analyzed on predefined set of standard datasets only, such as BOWS [Piva et al, 2007], BOSS [Bas et al, 2011] and ALASKA [Cogranne et al, 2019]. That's why it is of interest to performance analysis of CNN-based stegdetectors "in the wild" – on wide range of natural images.

Task and challenges

The paper is devoted to performance analysis of novel GB-Ras convolutional neural network based stegdetector "in the wild" by processing of natural images sampled from various open datasets. These datasets include digital images captured by various cameras as well as pre-processing histories (such as compression, filtering) that corresponds to SD working close to the real environment. The case of adaptive message hiding into cover image according to state-of-the-art methods S-UNIWARD and MG is considered.

Notations

High-dimensional arrays, matrices, and vectors will be typeset in boldface and their individual elements with the corresponding lower-case letters in italics. The symbols $\mathbf{U}=\left(u_{ij}\right)\in\Im^{N\times M}$, $\mathbf{X}=\left(x_{ij}\right)\in\Im^{N\times M}$ and $\mathbf{Y}=\left(y_{ij}\right)\in\Im^{N\times M}$, $\Im=\{0,1\ldots255\}$, will always represent pixel values of 8-bit grayscale initial (unprocessed), cover and stego images with size NxM pixels. The image's feature vector is denoted as $\mathbf{F}$, while the embedding binary message is represented as $\mathbf{M}$.

The Iverson bracket $[a]_I$ equals to one if the Boolean expression a is true, and zero otherwise. The notation $\|\cdot\|$ corresponds to Euclidean norm for scalar values, and Frobenius norm for matrices.

Digital images steganalysis with convolutional neural networks

The CNN-based architectures for supervised spatial image steganalysis was firstly proposed in 2015 by Qian et al. [Qian et al, 2015]. Proposed solution allowed achieving near state-of-the-art detection accuracy for wide range of embedding methods by preserving relatively low computation complexity of training procedure. This led to rapid growth of interest to applications of novel DL architectures in this domain, namely Xu-Net [Xu et al, 2016b], Ye-Net [Ye et al, 2017], Yedroudj-Net [Yedroudj et al, 2018], SR-Net [Boroumand et al, 2018], and Zhu-Net [Zhang et al, 2018]. The comparison of modern CNN-based stegdetectors architectures is presented at Fig. 2 [Reinel et al, 2021].

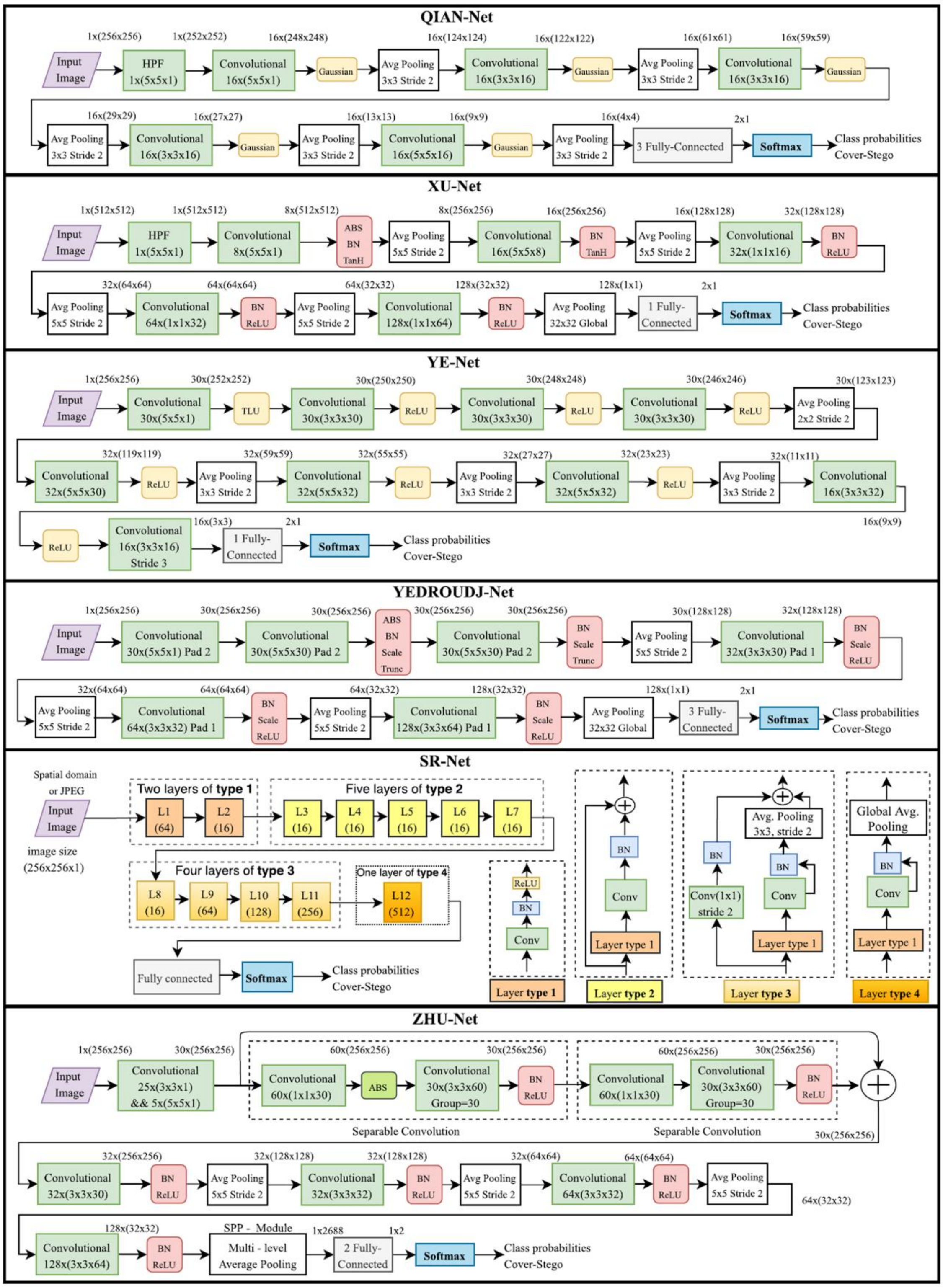


Figure 2. Comparison of architecture for modern CNN-based digital image stegdetectors [Reinel et al, 2021].

The seminal QIAN-Net uses set of high-pass filters to reduce DI content and emphasize steganographic noise. It was achieved by applying of convolutional layers with Gaussian activation function and Average Pooling for feature extraction. The classification of extracted features was performed by the final fully connected layer with softmax activation function.

The Xu-Net was proposed for additional improving detection accuracy of QIAN-Net [Xu et al, 2016b]. Feature of the network was usage of layer of absolute value as well as Batch Normalization to prevent SD overfitting. Further evolution of Xu-Net was Ye-Net proposed in 2017 [Ye et al, 2017] that included three improvements. Firstly, the model did not use the traditional high-pass filters to image content suppression. Instead, it used set of filters from SRM model [Fridrich et al, 2012] for initialization of network’s convolution kernels. Secondly, Truncated Linear Unit (TLU) activation functions are used in intermediate layers for increasing the Stego-to-Cover ratio. Finally, Ye-Net model incorporates methods for estimation probability of each pixel change (selection channel) for improving selection of appropriate convolutional kernels.

The Yedroudj-Net united mentioned features Xu-Net and Ye-Net networks [Yedroudj et al, 2018]. The convolutional layers of Yedroudj-Net were initialized with bank of 30 filters from SRM model, while TLU and batch normalization were used after all convolutions of feature extraction. Nevertheless, the Yedroudj-Net does not apply knowledge of selection channels that limits its performance for novel channel-aware embedding methods like Synch [Denemark et al, 2015].

The SR-Net network was proposed in 2018 for both spatial and JPEG domains steganlaysis [Boroumand et al, 2018]. Feature of the network is reducing usage of heuristic rules to capture the steganographic noise in comparison with considered networks. The SR-Net relies on a filter bank inspired by SRM to initialize weights in the pre-processing layers. Then, these weights may be updated during the training process to improving efficiency of image content suppression. The network has shown promising results for wide range of modern embedding methods and it is widely used today as reference for stegdetectors performance evaluation.

The Zhu-Net was aimed at reducing the size of the convolutional layer filters in comparison with SR-Net to decrease the number of parameters [Zhang et al, 2018]. Proposed separable convolutions enable estimation of residuals elements correlation and effective image content compression. The SPP [He et al, 2015] is used in the model to improving detection accuracy for various image sizes. The results obtained by Zhu-Net network outperformed obtained earlier by previous architectures.

Further evolution of Zhu-Net is GBRAS-Net model [Reinel et al, 2021] proposed in 2021. The network introduces several improvements in the stages of image pre-processing and feature extraction. Let us consider this model in more details (Fig. 3).

Digital image processing with GBRas-Net is done in several stages (Fig. 3). At the first stage, the inputted image is pre-processed (normalized) by convolution with 30 fixed kernels [Reinel et al, 2021]. In the paper [Reinel et al, 2021], it was proposed to initialize these kernels with corresponding filters form SRM model that allows obtaining high pre-processing effectiveness for subsequent feature extraction.

The next processing stage is aimed at feature extraction from pre-processed images. This is done by usage of trainable convolutional, separable convolutions, and depthwise convolutions layers. Also, the shortcuts are used at this stage for prevention of vanishing gradient. Then, extracted features are passed through dimensionality reduction part of model.

Finally, extracted features are used by classification part of GBRas-Net (Fig. 3). This part consists of global average pooling and further using of softmax activation functions for estimation probabilities to classify inputted image to cover or stego classes. It should be noted, dense layers are not used in classification stage that helps to avoid overfitting.

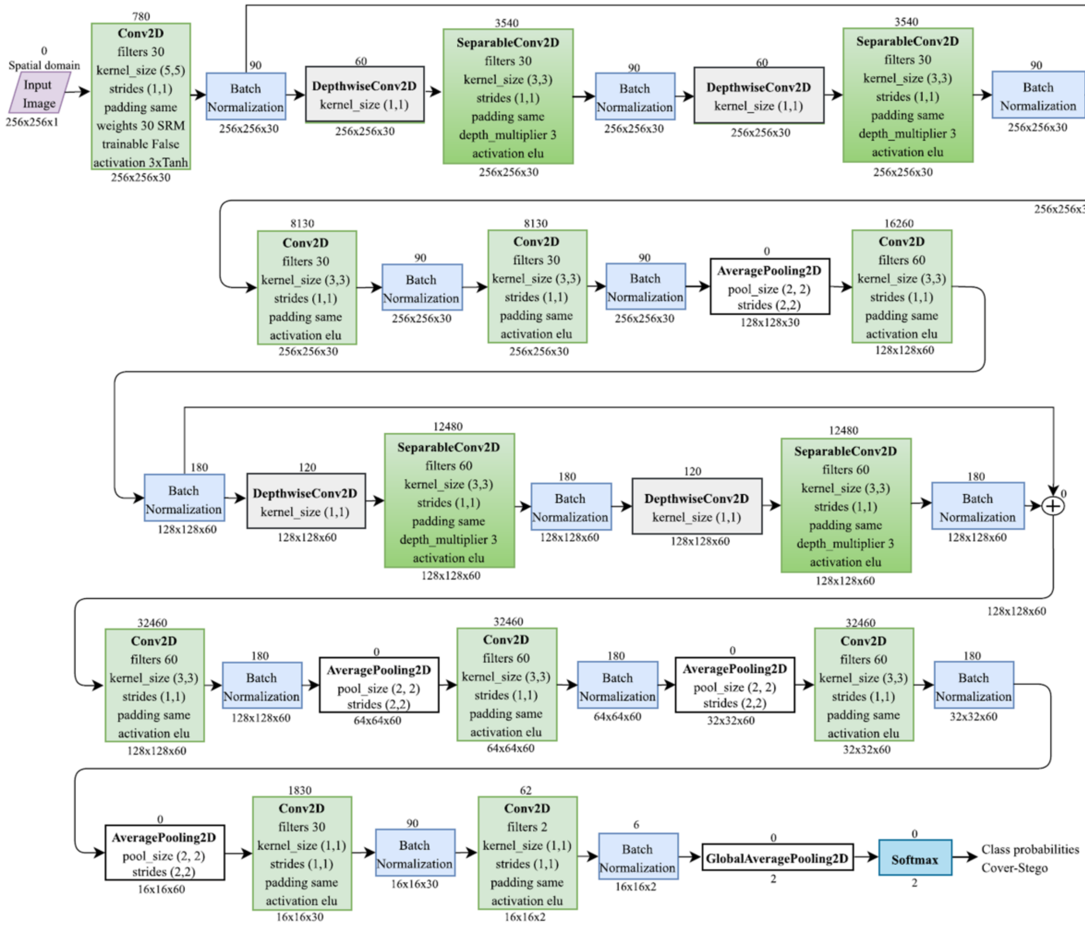


Figure 3. Architecture of GBRas-Net based stegdetector [Reinel et al, 2021].

The considered GBRas-Net provides several unique features that differentiate it from state-of-the-art CNN-based stegdetectors [Reinel et al, 2021]:

- The preprocessing stage in SR-Net is dropped and becomes an all-in-one with convolutional layers only, leaving the network to learn the convolutional kernels. Ye-Net and Zhu-Net take up the filter bank from SRM model. GBRAS-Net includes the bank of 30 filters from SRM model as non-trainable convolutional kernels in the image preprocessing stage.
- The GBRas allows achieving tradeoff between accuracy of residual estimation and complexity of theirs training with usage of 9 convolutional layers. In contrast, SR-Net provides high accuracy of residual estimation by the cost of high computation complexity of training the 25

convolutional layers, while Zhu-Net allows achieving fast training with 5 convolutional layers by the approximation of residuals.

- SR-Net and Zhu-Net have rectified linear unit activation functions in convolutional layers, while GBRAS-Net showed the best performance using the exponential linear unit activation function.
- SR-Net and Zhu-Net do not have depthwise convolutional Layers. GBRAS-Net has 4 such layers, which allow improving performance by detecting steganographic noise.

Mentioned features allows novel GBRas-Net provides high detection accuracy by preserving relatively low training complexity. The paper [Reinel et al, 2021] reported about improving detection accuracy for GBRas-Net in comparison with state-of-the-art CNN-based stegdetectors for adaptive embedding methods S-UNIWARD and WOW on standard BOWS, BOSS and ALASKA datasets.

Adaptive embedding methods

The majority of state-of-the-art embedding methods are based on minimizing the empirical distortion estimation function $D(\mathbf{X},\mathbf{Y})$ during forming a stego image $\mathbf{Y}$ from a cover image (CI) $\mathbf{X}$ [Filler et al, 2011]:

$$D(\mathbf{X},\mathbf{Y}) = \sum_i \rho_i(\mathbf{X},\mathbf{Y}) \to \min, |\mathbf{M}| = const, \tag{1}$$

where $\rho_i(\cdot)$ – function for estimation the alteration of cover image's statistical characteristics by embedding of ith stegobit; $|\mathbf{M}|$ – size of embedded message $\mathbf{M}$, bits. Minimization of function (1) allows adapting the embedding process to a cover image, thus corresponding steganographic methods called adaptive.

In most cases, the choice of function $D(\mathbf{X},\mathbf{Y})$ is done under the assumption of independency of distortions caused by embedding of individual stegobits (distortions additivity) that simplifies the choice of function (1). Nevertheless, this approach does not taking into account interactions between distortions that may lead to non-linear changes of CI parameters.

In this paper we considered the case of usage the state-of-the-art adaptive embedding methods, namely S-UNIWARD [Holub et al, 2014] and MG [Sedighi et al, 2015]. These methods are aimed on message in spatial domain of CI with size $N \times M$ pixels – by manipulation with brightness of individual pixels. Let us consider these methods in details.

The S-UNIWARD embedding method is based on spectral transformation of CI. The transformation is used for estimation the cover image distortions caused by embedding of individual stegobits. Similarly to HUGO method, S-UNIWARD method takes additive empirical distortion estimation function [Holub et al, 2014]:

$$D(\mathbf{X},\mathbf{Y}) = \sum\nolimits_{k,u,v} \frac{\left|\mathbf{W}_{uv}(\mathbf{X},k) - \mathbf{W}_{uv}(\mathbf{Y},k)\right|}{\sigma + \left|\mathbf{W}_{uv}(\mathbf{X},k)\right|}, \tag{2}$$

where $\mathbf{W}_{uv}(\mathbf{X},k)$, $\mathbf{W}_{uv}(\mathbf{Y},k)$ – coefficients of two-dimensional discrete wavelet transform (2D-DWT) of the cover $\mathbf{X}$ and stego $\mathbf{Y}$ images with coordinates (u,v) in the k^{th} sub-band; $\sigma > 0$ – stabilizing constant.

Variation of 2D-DWT basis functions in (2) allows analyzing specific distortions of CI caused by message hiding. Also, usage of empirical distortion estimation function (2) makes it possible to message hiding in spatial (alteration of CI pixels brightness) and transformation (by changing of a CI decomposition coefficients) domains in the uniform way.

The alternative approach to design an empirical distortion estimation function (1) is based on minimization both cover image distortions and statistical detectability of formed stego images [Ker et al, 2013]. As an example of such embedding method, the MG [Sedighi et al, 2015] and MiPOD [Sedighi et al, 2016] embedding methods can be taken. Feature of these methods is usage of locally-estimated multivariate Gaussian cover image model. It allows achieving the stego image’s empirical security that is comparable with advanced steganographic methods.

Formation of stego images according to MiPOD method is carried out in several steps [Sedighi et al, 2016]. Firstly, it is suppressed the cover image context $\mathbf{X}=\left(x_1,\ldots,x_{M\cdot N}\right)$, using denoising filter F :

$$\mathbf{r}=\mathbf{X}-F\left(\mathbf{X}\right),$$

where $\mathbf{X}$ is represented in column-wise order. Then, it is measured pixels residual variance σ_l^2 using Maximum Likelihood Estimation:

$$\mathbf{r}_l=\mathbf{G}\mathbf{a}_l+\boldsymbol{\xi}_l, \tag{3}$$

where $\mathbf{r}_l -$ represents the value of the residuals $\mathbf{r}$ inside the $p\times p$ block surrounding the lth residual put into a column vector; $\mathbf{G}_{p^2\times p} -$ the matrix defines the parametric model of remaining expectation; $\mathbf{a}_{p\times 1} -$ the vector of linear model's parameters; $\boldsymbol{\xi}_{p^2\times 1} -$ the signal whose variance is need to be estimated.

At the second step, the pixels residual variance σ_l^2 is estimated according to formula:

$$\sigma_l^2=\left\|\mathbf{P}_{\mathbf{G}}^{\perp}\mathbf{r}_l\right\|^2\Big/\left(p^2-q\right),$$

where $\mathbf{P}_{\mathbf{G}}^{\perp}=\mathbf{I}_l-\mathbf{G}\left(\mathbf{G}^T\mathbf{G}\right)^{-1}\mathbf{G}^T -$ the orthogonal projection of residual $\mathbf{r}_l$ (3) to the p^2-q dimensional subspace spanned by the left null space of $\mathbf{G}$; $\mathbf{I}_{l\times l} -$ the unity matrix.

The MG embedding method [Sedighi et al, 2015] is similar to MiPOD algorithm, but it uses the simplified variance estimator:

$$\sigma_l^2=\left\|\mathbf{r}_n-\hat{\mathbf{r}}_n\right\|^2\Big/\left(p^2-q\right),$$

$$\hat{\mathbf{r}}_n = \mathbf{G}\left(\mathbf{G}^T\mathbf{G}\right)^{-1}\mathbf{G}^T\mathbf{r}_n.$$

Thirdly, it is determined the probability of lth embedding change $\beta_l, l \in \{1,2,\ldots,L\}$ that minimize the deflection coefficient ς^2 between cover and stego image distributions:

$$\zeta^2 = 2\sum_{l=1}^{M \cdot N} \beta_l^2 \sigma_l^{-4}, \tag{4}$$

under payload constrain

$$R = \sum_{l=1}^{M \cdot N} H\left(\beta_l\right),$$

where $H(z) = -2z\log z - (1-2z)\log(1-2z) -$ ternary entropy function; $R -$ cover image payload in nats.

Minimization of (4) can be achieved by using the method of Lagrange multipliers. The change rate β_l and the Lagrange multiplier λ can be determined by numerically solving of next $(l+1)$ equations:

$$\beta_l \sigma_l^{-4} = \frac{1}{2\lambda}\ln\left(\frac{1-2\beta_l}{\beta_l}\right), l \in [1; M \cdot N],$$

$$R = \sum_{l=1}^{M \cdot N} H\left(\beta_l\right).$$

Then, the change rate β_l is converted to the cost ρ_l:

$$\rho_l = \ln\left(1/\beta_l - 2\right). \tag{5}$$

Finally, the desired payload R is embedded using syndrome-trellis codes (STCs) with pixel costs determined according to (5).

Applying the locally-estimated multivariate Gaussian cover model in MiPOD algorithm gives opportunity to derive a closed-form expression for the performance of the stegdetector and capture the non-stationary character of natural images [Sedighi et al, 2016].

Experiments

Performance analysis of modern GB-Ras network based stegdetector was done for state-of-the-art adaptive S-UNIWARD and MG embedding methods. Cover image payload was was changed within range 3%, 5%, 10%, 20%, 30%, 40% and 50%. Evaluation was performed for standard ALASKA dataset [Cogranne et al, 2019] as well as publicly available datasets of natural images VISION [Shullani et al, 2017] and MIRFlickr [Huiskes et al, 2008]:

- The ALASKA dataset [Cogranne et al, 2019] is widely used today for performance evaluation of modern SD. It consists of 80,000 images captured by 40 cameras, including smartphones, tablets, low-end cameras to high-end full frame digital single-lens reflex camera (DLSR). The images were processed from raw files in a realistic and highly heterogeneous way, namely by variation of demosaicing algorithm, resizing, image compression methods etc.
- The VISION dataset [Shullani et al, 2017] was proposed in DI forensics domain, namely for noise parameters estimation and camera identification. The dataset consists of 34,427 images and 1,914 videos, both in the native format and in their social version (Facebook, YouTube, and WhatsApp are considered). The data were collected by 35 mobile devices (smartphones) of 11 major brands, such as Apple, Samsung, Huawei, LG, Sony etc.
- The MIRFlickr dataset [Huiskes et al, 2008] is part of open evaluation project of image forensics and image retrieval. The dataset includes DI from the photography service Flickr coupled with complete manual annotations, pre-computed descriptors and EXIF data (metadata about camera settings during image capturing). The variety of image sizes and pre-processing stages (such as compression, visual enhancement) makes MIRFlickr dataset an attractive candidate for performance

evaluation of SD "in the wild" – in the environment as much close as possible to the real cases of natural images processing.

During research, the sub-sets of 10,000 DI were sampled from each of mentioned datasets. The images were converted to grayscale format (if they were colored) with usage of standard funciton "rgb2gray" from MATLAB programming and numeric computing platform. Finally, images were cropped to preserve similar size of 512x512 pixels.

The SD was based on usage the pre-trained GB-Ras models [Reinel et al, 2021]. The model was trained for S-UNIWARD and WOW methods on BOSS dataset [Bas et al, 2011] by 20% and 40% cover image payload. Since pre-trained networks require 256x256 inputs, prepared test images were divided into 256x256 tiles (4 tiles per image). The decision about classifying an image as cover or stego one was performed by majority voting on each tiles related to the image.

For estimation baseline detection accuracy, the statistical SD based on modern SPAM [Pevny et al, 2010] and maxSRMd2 [Denemark et al, 2014] cover rich models were used. The SPAM model is based on applying of first and second order Markov chain for adjacent pixels brightness correlation modeling. The maxSRMd2 model relates to the group of SRM models [Fridrich et al, 2012]. The model is based on image pre-processing with extensive set of high-pass filters for cover's content suppression, and further applying of Markov chains for modeling dependencies between adjacent elements of extracted residuals [Denemark et al, 2014]. Both stegdetectors were trained on stego images formed according to S-UNIWARD embedding method on ALASKA dataset for providing similar evaluation pipeline for GB-Ras network. Therefore, the ALASKA dataset was pseudorandomly splitted into two parts with equal sizes – the first part was used for statistical SD training, while the second was used for stegdetector evaluation. Then, pre-trained SD was used for evaluation of SPAM and maxSRMd2 model performance by processing images from VISION and MIRFlickr datasets.

The total classification error P_E was used as metric during SD performance evaluation [Kodovsky et al, 2012]:

$$P_E = \min_{P_{FA}} \frac{1}{2}\left(P_{FA} + P_{MD}\left(P_{FA}\right)\right),$$

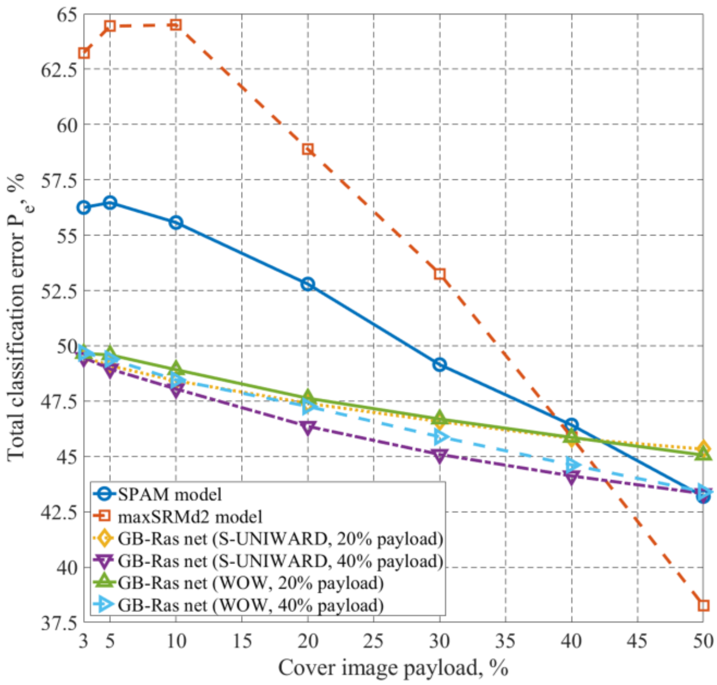
where P_{FA} and P_{MD} are probability of false alarm and missed detection, respectively.

Investigation of GB-Ras network performance was done in several stages by using of mentioned datasets. At the first stage, the case of using the standard ALASKA dataset was considered. The dependencies of total error P_E on CI payload for S-UNIWARD and MG embedding methods by the usage of GB-Ras and cover rich models based stegdetector on ALASKA dataset are presented at Fig.4.

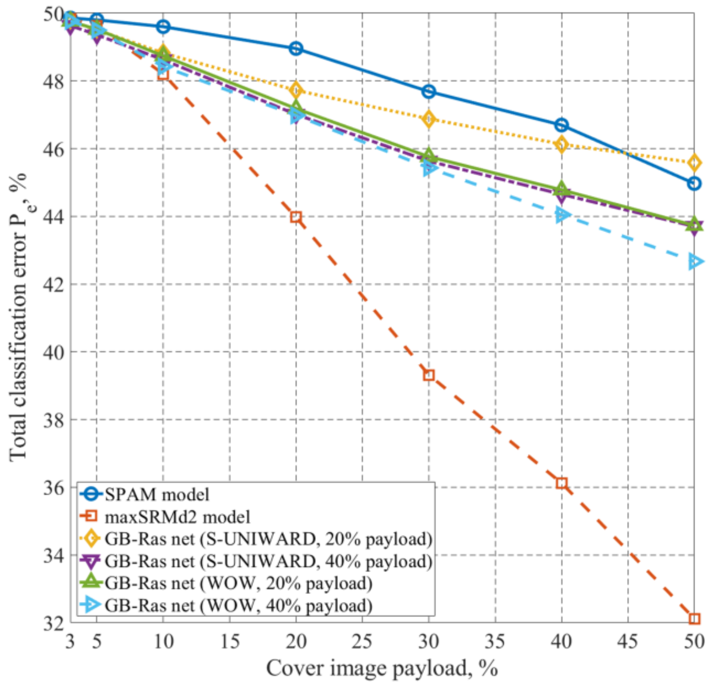
It should be noted considerable decreasing of detection error by applying of GB-Ras based SD for S-UNIWARD embedding method (Fig. 4a) – applying of such stegdetector allows decreasing P_E up to 6% for the most difficult case of low CI payload (less than 10%). The mentioned decreasing is preserved even for medium cover image payload (Fig. 4a, less than 25%), where GB-Ras allows to outperform the state-of-the-art maxSRMd2 mode up to 8%. Nevertheless, for the high payload (more than 30%) the cover rich model based SD allows achieving smaller detection error by the cost of higher computation complexity.

On the other hand, GB-Ras based SD is conceded to maxSRMd2 stegdetector for MG embedding method (Fig. 4b) – they performed similarly for the low CI payload, while rich model based stegdetector considerable improved detection accuracy by increasing of payload. Nevertheless, the GB-Ras outperformed the standard SPAM model based SD for all cover image payload up to 3%.

For comparison, the dependencies of total error P_E on CI payload for S-UNIWARD and MG embedding methods by the usage of GB-Ras and cover rich models based stegdetector on VISION dataset are presented at Fig.5.

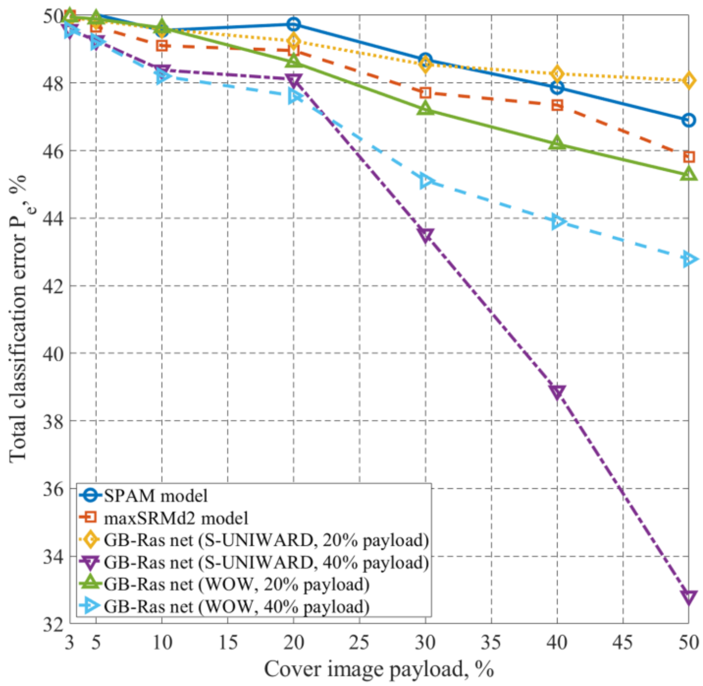


a)

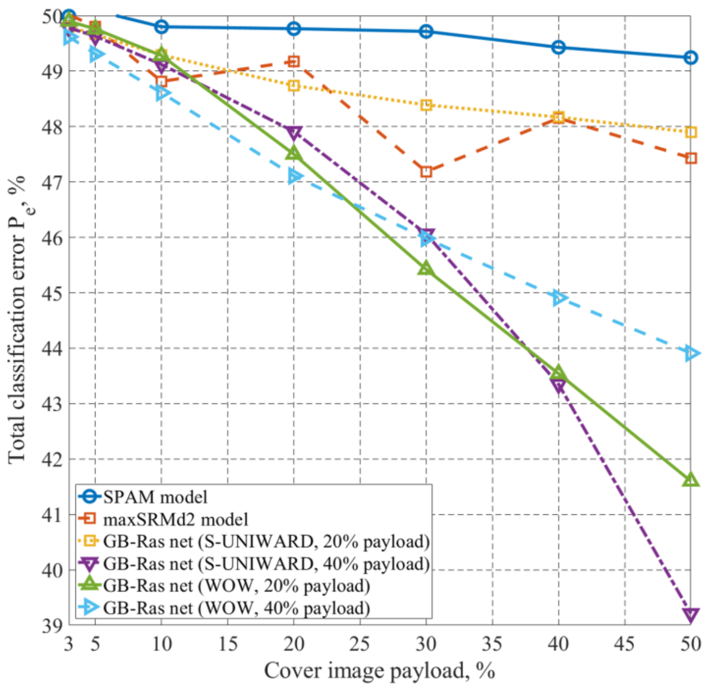


b)

Figure 4. The dependencies of total error P_E on cover image payload for S-UNIWARD (a) and MG (b) embedding methods by the usage of GB-Ras and cover rich models based stegdetector on ALASKA dataset. The options of pre-trained GB-Ras network are presented in the parentheses.



a)



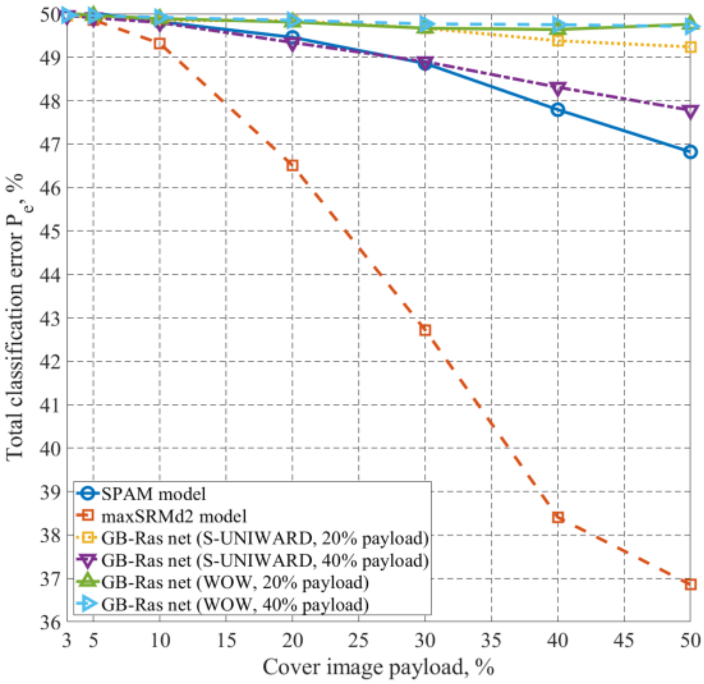
b)

Figure 5. The dependencies of total error P_E on CI payload for S-UNIWARD (a) and MG (b) embedding methods by the usage of GB-Ras as well as cover rich models based stegdetector on VISION dataset. The pre-training conditions for GB-Ras network are presented in the parentheses.

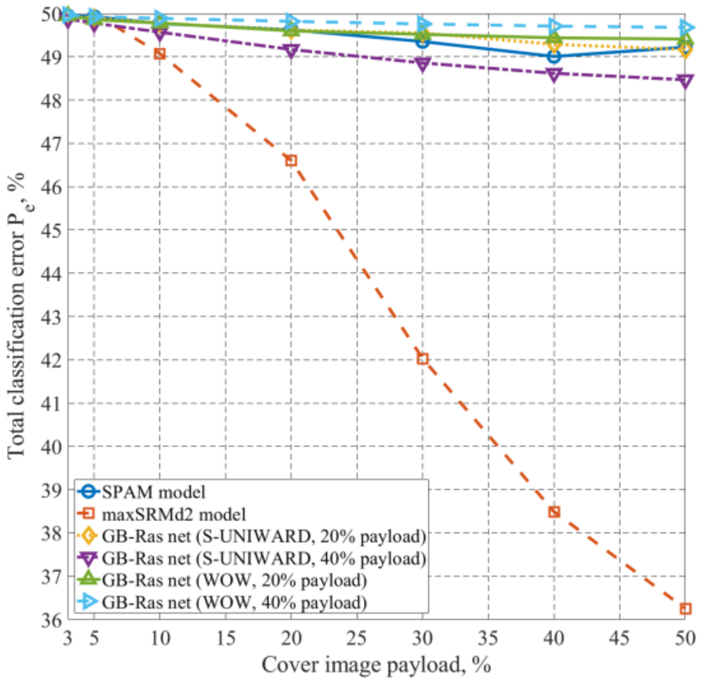
In contrast to the previous case (Fig .4), applying of GB-Ras based stegdetector allows achieving similar or even smaller detection error for both S-UNIWARD and MG embedding methods on VISION dataset in comparison with rich model based SD (Fig. 5). In this case, considered stegdetectors performed with similar detection error in the low CI payload range (less than 10%), while accuracy of GB-Ras increased drastically by increasing of the payload (Fig .5) – up to 13% for S-UNIWARD (Fig. 5a) and to 9% for MG (Fig. 5b) embedding methods in comparison with RM based stegdetectors. Obtained results may be explained by quite high noise level of test images from VISION datasets in comparison with other datasets [Shullani et al, 2017]. Therefore, usage of pre-trained convolutional kernels for GB-Ras model allows effectively suppressing these noises without necessity to applying of extensive set of high-pass filters for cover rich models.

Finally, the dependencies of total error P_E on CI payload for S-UNIWARD and MG embedding methods by the usage of GB-Ras and cover rich models based stegdetector on MIRFlickr dataset are represented at Fig.6.

The maxSRMd2 based stegdetector allows considerably decreasing detection error for images sampled from MIRflickr dataset in comparison with GB-Ras network (Fig. 6) – the differences are up to 11% for S-UNIWARD and 12% for MG embedding methods. The SPAM model based SD performed similar to GB-Ras or even worse for the case of high cover image payload (more than 30%). Obtained results may be explained by high quality of DI from MIRflickr dataset, namely low noise level. Therefore, maxSRMd2 model allows capturing more information about cover image content by applying of extensive set of high-pass filters in comparison with GB-Ras model.



a)



b)

Figure 6. The dependencies of total error P_E on CI payload for S-UNIWARD (a) and MG (b) embedding methods by the usage of GB-Ras as well as cover rich models based stegdetector on MIRflickr dataset. The pre-training conditions for GB-Ras network are presented in the parentheses.

Conclusion

In the paper we argue that usage of pre-trained convolutional neural network allows decreasing negative impact of domain adaptation problem for modern stegdetectors. The case of stego image detection by novel GB-Ras and state-of-the-art cover rich models stegdetectors was considered for advanced S-UNIWARD and MG embedding methods.

The GB-Ras model allows improving detection accuracy of stego images for standard ALASKA dataset for S-UNIWARD embedding methods (up to 6% for the most difficult case of low cover image payload). At the same time, the model gives opportunity to considerably decreasing of detection error for VISION dataset of natural images collected by mobile devices – up to 13% for S-UNIWARD (Fig. 5a) and to 9% for MG (Fig. 5b) embedding methods in comparison with RM based stegdetectors. Nevertheless, the performance of considered GB-Ras based stegdetector for high quality images from MIRFlickr dataset is conceded to maxSRMd2 rich model based stegdetector (up to 11% for S-UNIWARD and 12% for MG embedding methods).

Obtained results proved the effectiveness of neural network applying for mitigation with domain adaptation problem for modern stegdetectors. The considered GB-Ras network may be used as a promising candidate for further design of novel stegdetectors that are robust to domain adaptation problem.

Bibliography

[Bas et al, 2011] Bas P., Filler T., Pevný T. "Break Our Steganographic System": The Ins and Outs of Organizing BOSS. International Workshop on Information Hiding, 2011. DOI: 10.1007/978-3-642-24178-9_5.

[Bisong, 2019] Bisong E. Regularization for Deep Learning. Building Machine Learning and Deep Learning Models on Google Cloud Platform, pp.415-421, 2019. DOI:10.1007/978-1-4842-4470-8_34

[Boroumand et al, 2018] Boroumand M., Chen M., Fridrich J. Deep Residual Network for Steganalysis of Digital Images. IEEE Trans. Inf. Forensics Security, vol. 14, issue 5, 2018, pp. 1181-1193.

[Cogranne et al, 2019] Cogranne R., Giboulot Q., Bas P. The ALASKA Steganalysis Challenge: A First Step Towards Steganalysis. ACM Workshop on Information Hiding and Multimedia Security, pp. 125-137, 2019. DOI: https://doi.org/10.1145/3335203.3335726.

[Denemark et al, 2014] Denemark T., Sedighi V., Holub V., Cogranne R., Fridrich J. Selection-Channel-Aware Rich Model for Steganalysis of Digital Images. IEEE Workshop on Information Forensic and Security, Atlanta, GA, 2014. DOI: 10.1109/WIFS.2014.7084302

[Denemark et al, 2015] Denemark T., Fridrich J. Improving Steganographic Security by Synchronizing the Selection Channel. 3rd IH&MMSec. Workshop, Portland, Oregon, June 17-19, 2015, DOI: 10.1145/2756601.2756620.

[Filler et al, 2011] Filler T., Fridrich J. Design of Adaptive Steganographic Schemes for Digital Images. Proc. SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics XIII, 2011, DOI: 10.1117/12.872192.

[Fridrich et al, 2012] Fridrich J., Kodovsky J. Rich Models for Steganalysis of Digital Images. IEEE Trans. Inf. Forensics Security, vol. 7, issue 3, 2012, pp. 868-882.

[Fridrich, 2009] J. Fridrich Steganography in Digital Media: Principles, Algorithms, and Applications. 1st edition. Cambridge University Press, 2009. p. 437. ISBN 978–0–521–19019–0;

[He et al, 2015] He K., Zhang X., Ren S., Sun J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904- 1916, 2015.

[Holub et al, 2014] Holub V., Fridrich J., Denemark T. Universal Distortion Function for Steganography in an Arbitrary Domain. EURASIP Journal on Information Security, Vol. 1, 2014, DOI: 10.1186/1687-417X-2014-1.

[Huiskes et al, 2008] Huiskes M. J., Lew M.S. The MIR Flickr Retrieval Evaluation. ACM International Conference on Multimedia Information Retrieval, Canada, 2008. DOI: 10.1145/1460096.1460104

[Jin et al, 2020] Jin Z., Yu. Y, Chen Y., Chen Y. IAS-CNN: Image adaptive steganalysis via convolutional neural network combined with selection channel. International Journal of Distributed Sensor Networks, Vol. 16 iss. 3, Mar. 2020, DOI: 10.1177/1550147720911002

[Ker et al, 2013] Ker A. D., Bas P., Böhme R., Cogranne R., Craver S., Filler T., Fridrich J., Pevný T. Moving steganography and steganalysis from the laboratory into the real world. Proceedings of the first ACM workshop on Information hiding and multimedia security (IH&MMSec '13). New York, 2013;

[Kodovsky et al, 2012] Kodovsky J., Fridrich J., Holub V. Ensemble Classifiers for Steganalysis of Digital Media. IEEE Trans. Inf. Forensics Security, vol. 7, issue 2, 2012, pp. 432-444.

[Konahovych et al, 2018] Konachovych G., Progonov D., Puzyrenko O. Digital steganography processing and analysis of multimedia files [In Ukrainian]. ‘Tsentr uchbovoi literatury’ publishing, 2018, ISBN 978-617-673-741-4.

[Kopeytsev, 2020] Kopeytsev V. Steganograph in attacks on industrial enterprises. Kaspersky Inc., Tech. Rep, 2020, 6 pages. [Online] Available :https://ics-cert.kaspersky.com/media/KASPERSKY_Steganography_in_targeted_attacks_EN.pdf

[Pevny et al, 2010] Pevny T., Bas P., Fridrich J. Steganalysis by Subtractive Pixel Adjacency Matrix. IEEE Trans. Inf. Forensics Security, vol. 5, issue 2, 2010, pp. 215-224.

[Piva et al, 2007] Piva, A., Barni, M. The first BOWS contest: Break our watermarking system. In: Delp, E.J., Wong, P.W. (eds.) Proceedings SPIE, Electronic Imaging, Security, Steganography, and Watermarking of Multimedia Contents IX. vol. 6505. San Jose, CA, 2007.

[Qian et al, 2015] Qian Y., Dong J., Wang W., Tan T. Deep learning for steganalysis via convolutional neural networks. Proc. IS T Int. Symp. Electron. Imag. (EI), vol. 9409, 2015.

[Reinel et al, 2019] Reinel T., Raúl R., Gustavo I. Deep Learning Applied to Steganalysis of Digital Images: A Systematic Review. IEEE Press, 2019, Vol.7. pp. 68970 – 68990, DOI: 10.1109/ACCESS.2019.2918086.

[Reinel et al, 2021] Reinel T.-S., Brayan A.-A. H., Alejandro B.-O. M., Alejandro M.-R., Daniel A.-G., Alejandro A.-G. J. Buenaventura B.-J. A. Simon O.-A. Gustavo I., Raúl R.-P. GBRAS-Net: A Convolutional Neural Network Architecture for Spatial Image Steganalysis. IEEE Access, Vol. 9, pp. 14340 – 14350, 2021, DOI: 10.1109/ACCESS.2021.3052494.

[Sedighi et al, 2015] Sedighi V., Fridrich J., Cogranne R. Content-Adaptive Pentary Steganography Using the Multivariate Generalized Gaussian Cover Model. Proc. SPIE, Electronic Imaging, Media Watermarking, Security, and Forensics, 2015, vol. 9409.

[Sedighi et al, 2016] Sedighi V., Cogranne R., Fridrich J. Content-Adaptive Steganography by Minimizing Statistical Detectability. IEEE Trans. Inf. Forensics Security. Vol. 11, Iss. 2., 2016. pp. 221-234.

[Shullani et al, 2017] Shullani D., Fontani M., Iuliani M., Al Shaya O., Piva A. VISION: a video and image dataset for source identification. EURASIP Journal on Information Security, Vol. 15, 2017. DOI: 10.1186/s13635-017-0067-2.

[Singh et al, 2019] Singh P., Verma V. K., Rai P. HetConv: heterogeneous kernel-based convolutions for deep CNNs. 2019, Available: http://arxiv.org/abs/1903.04120

[Tan et al, 2014] Tan S., Li B. Stacked convolutional auto-encoders for steganalysis of digital images. Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific, pages 1–4, December 2014, DOI: 10.1109/APSIPA.2014.7041565.

[Xu et al, 2016a] Xu G., Wu H.-Z., Shi Y. Q. Ensemble of CNNs for steganalysis: An empirical study. 4th ACM Workshop Inf. Hiding Multimedia Secur., 2016, pp. 103-107.

[Xu et al, 2016b] Xu G., Wu H.-Z., Shi Y.-Q. Structural design of convolutional neural networks for steganalysis. IEEE Signal Process. Lett., vol. 23, no. 5, pp. 708-712, 2016.

[Xu et al, 2020] Xu Y., Fu Z., Xu G., Zhang S., Xie X. DRHNet: A Deep Residual Network Based on Heterogeneous Kernel for Steganalysis. Secur. Commun. Networks, Vol. 20, 2020. DOI: 10.1155/2020/8847741

[Yang et al, 2018] Yang J., Liu K., Kang X., Wong E. K., Shi Y.-Q. 'Spatial image steganography based on generative adversarial network. arXiv:1804.07939, 2018. [Online]. Available: https://arxiv.org/abs/1804.07939

[Ye et al, 2017] Ye J., Ni J., Yi Y. Deep Learning Hierarchical Representations for Image Steganalysis. IEEE Transactions on Information Forensics and Security, vol. 12, no. 11, pp. 2545-2557, 2017. DOI: 10.1109/TIFS.2017.2710946.

[Yedroudj et al, 2018] Yedroudj M., Comby F., Chaumont M. Yedroudj-Net: An efficient CNN for spatial steganalysis. Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2018, pp. 2092-2096.

[Zhang et al, 2018] Zhang R., Zhu F., Liu J., Liu G. 'Efficient feature learning and multisize image steganalysis based on CNN. arXiv:1807.11428, 2018. [Online]. Available: http://arxiv.org/abs/1807.11428.

Authors' Information

Dmytro Progonov – *Institute of Physics and Technology, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"; PhD, Associate Professor; 37, ave. Peremohy, Solomenskiy district, Kyiv, Postcode 03056, Ukraine;* ***e-mail****: progonov@gmail.com*

Major Fields of Scientific Research*: digital media steganalysis, digital image forensics, machine learning, advanced signal processing*